JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Promiscuous Histone Mis-Assembly Is Actively Prevented by Chaperones

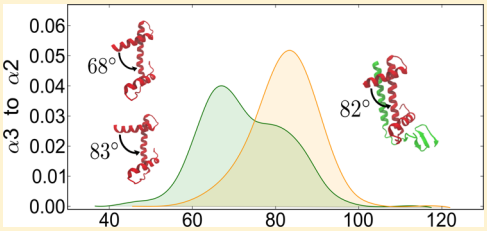
Haiqing Zhao,[†,‡,∥] David Winogradoff,[⊥,∥] Minh Bui,[‡] Yamini Dalal,[*,‡] and Garegin A. Papoian[*,†,⊥,§]

[†]Biophysics Program, [⊥]Chemical Physics Program, and [§]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

Supporting Information

ABSTRACT: Histone proteins are essential for the organization, expression, and inheritance of genetic material for eukaryotic cells. A centromere-specific H3 histone variant, centromere protein A (CENP-A), shares about 50% amino acid sequence identity with H3. CENP-A is required for packaging the centromere and for the proper separation of chromosomes during mitosis. Despite their distinct biological functions, previously reported crystal structures of the CENP-A/H4 and H3/H4 dimers reveal a high degree of similarity. In this work, we characterize the structural dynamics of CENP-A/H4 and H3/H4 dimers based on a dual-resolution approach, using both microsecond-scale explicit-solvent all-atom and coarse-grained (CG) molecular dynamics (MD) simulations. Our data show that the H4 histone is significantly more rigid compared with the H3 histone and its variant CENP-A, hence, serving as a reinforcing structural element within the histone core. We report that the CENP-A/H4 dimer is significantly more dynamic than its canonical counterpart H3/H4, and our results provide a physical explanation for this flexibility. Further, we observe that the centromere-specific chaperone Holliday Junction Recognition Protein (HJURP) stabilizes the CENP-A/H4 dimer by forming a specific electrostatic interaction network. Finally, replacing CENP-A S68 with E68 disrupts the binding interface between CENP-A and HJURP in all-atom MD simulation, and consistently, *in vivo* experiments demonstrate that replacing CENP-A S68 with E68 disrupts CENP-A's localization to the centromere. Based on all our results, we propose that, during the CENP-A/H4 deposition process, the chaperone HJURP protects various substructures of the dimer, serving both as a folding and binding chaperone.



INTRODUCTION

In eukaryotes, genomic DNA associates with histone proteins, assembling into arrays of nucleosomes. The canonical nucleosome contains 147 base pairs of DNA, wrapped around the histone octamer core with two copies each of the histones H2A, H2B, H3, and H4.[1] These core histones are among the most conserved proteins in eukaryotes, and all feature the same structural motif, known as the "histone-fold."[2] However, recent studies revealed that variant histones have evolved for diverse and specific functions.[3−7] Extensive studies in cell biology, biochemistry, and biophysics have interrogated the relationships between the sequence, structure, and function of histone variants in various biological contexts.[3−9] Indeed, variation in histone primary sequence serves as the foundation of genomic regulation *in vivo* by leading to functional changes in chromatin structure and dynamics.[10,11] In contrast to all the other core histones, there are no reported variants of H4.[12] Whether the absence of histone variants for H4 reflects greater structural integrity remains unknown, and addressing this question may shed light on the structural foundation of genetic inheritance.

Within the H3 family, the variant CENP-A (CenH3) specifies the unique location of the centromere required for proper chromosome segregation during cell division. In particular, CENP-A is reported to be overexpressed and mislocalized into noncentromeric chromosome regions in aggressive cancer cells.[13,14] Interestingly, the crystal structures of CENP-A and canonical H3 are nearly identical, except for minor differences in CENP-A's αN helix, and loop 1 regions.[15,16] However, *in vivo* CENP-A-containing nucleosomes have been shown to occupy a multitude of structures.[17−34] Our recent all-atom molecular dynamics (MD) study revealed that the octameric CENP-A nucleosome displays more structural heterogeneity on a local and global scale than its H3 counterpart,[35] a result that has since been experimentally validated by FRET assays demonstrating that CENP-A octameric nucleosomes *in vitro* are highly flexible,[36] in contrast to previous reports that the CENP-A nucleosome is rigidified[25,37] *in vitro*. Since the CENP-A dimer is the key component distinguishing the CENP-A nucleosome from the canonical H3 nucleosome, we were curious whether, in isolation or coupled to its chaperone Holliday junction recognition protein (HJURP), the CENP-A/H4 dimer displays

Received: May 25, 2016
Published: July 25, 2016

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

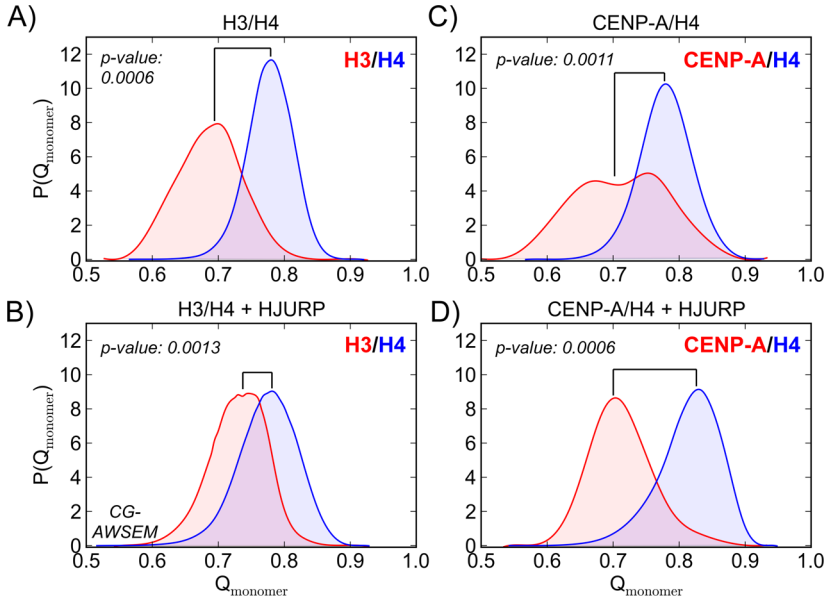


Figure 1. H4 adopts conformations closer to the native state than CENP-A or H3 in CG-AWSEM simulations. $Q_{monomer}$ characterizes a monomer's structural resemblance to its native state, defined by the corresponding monomeric conformations found in the crystal structures for H3/H4 (PDB ID: 1AOI)[1] and CENP-A/H4 (PDB ID: 3R45).[16] Probability distributions of monomer Q are plotted for either H3 vs H4 or CENP-A vs H4 in (A) the H3/H4 dimer, (B) the CENP-A/H4 dimer, (C) H3/H4 in the presence of HJURP, and (D) the CENP-A/H4/HJURP complex. For each system, the average monomer Q value for H4 (blue) is greater than the average for CENP-A or H3 (red). Matching the CG-AWSEM results, H4 is structurally consistent in all-atom MD simulations (Figure S2).

dynamics distinct from that of H3/H4, which might, in turn contribute to its unique biology *in vivo*.

Investigating the dynamics of histone variant deposition into and eviction from nucleosomes is fundamentally important, with chaperones like HJURP playing a key role in facilitating and regulating histone delivery, exchange, and removal.[38,39] The chaperone HJURP has been demonstrated to be required for the deposition of CENP-A into the kinetochore,[40−42] but precisely how HJURP dynamically interacts with CENP-A/H4 and how HJURP mediates CENP-A's deposition through these interactions remain unclear.

To address the questions above, one could rely on molecular simulations of the CENP-A/H4 and H3/H4 dimers and also the ternary complexes with HJURP. Usually, either atomistic or coarse-grained simulations are chosen for such studies, where the former provides finer resolution but samples less conformational space, raising issues of convergence for systems of this size. Coarse-grained simulations, on the other hand, quickly achieve equilibration, however, detailed atom-by-atom structural interactions are averaged over. In this work, we studied the same systems employing a novel dual-resolution approach, using both coarse-grained AWSEM[43] (CG-AWSEM) and all-atom molecular dynamics (MD) simulations. These two techniques complement each other: CG-AWSEM MD (i.e., three beads per amino acid residue) in implicit solvent samples more conformational space and explores more global properties of the histone dimers, whereas all-atom MD in explicit solvent probes specific interactions and native-state dynamics at high resolution. One of the overarching goals of our work was to cross-validate the conclusions obtained from these two independent methods, analyzing consistent findings or discrepancies in some detail.

Both CG-AWSEM and all-atom results indicate that histone H4 adopts configurations closer to the native state than either CENP-A or H3, demonstrating the structural resilience that is predicted from its high sequence conservation and the absence of variants. The CENP-A/H4 dimer is more structurally variable than the canonical H3/H4 dimer in CG-AWSEM simulations, wherein the dimer interface of CENP-A/H4, in particular, exhibits greater conformational heterogeneity. A key component that distinguishes the dynamics of CENP-A/H4 from H3/H4 is the longer and more acidic C-terminal residues of CENP-A, which, in our simulation results, is surprisingly regulated by its chaperone HJURP. In all-atom MD simulations, we observe that HJURP facilitates the formation of a structure-inducing electrostatic network with the C-termini of CENP-A and H4 and that the N-terminal portion of CENP-A containing S68 forms key interactions with a hydrophobic pocket of HJURP. To test the hypothesis that CENP-A S68 is required for binding with HJURP, we performed *in vivo* experiments and all-atom simulations mutating this residue. Finally, we discuss the implications of our findings on the recruitment of other centromeric proteins, such as CENP-C, and propose a model in which HJURP may play dual roles in guiding CENP-A's deposition, serving both as a folding and a binding chaperone.

RESULTS

In this work, we performed microsecond-scale coarse-grained and explicit-solvent atomistic MD simulations for the following systems: (1) the H3/H4 dimer; (2) the CENP-A/H4 dimer; (3) the CENP-A/H4/HJURP complex; (4) the H3/H4 dimer with HJURP. Initial conformations are based on the crystal structures of the canonical nucleosome (PDB ID: 1AOI)[1] and of the CENP-A/H4 dimer with chaperone HJURP (PDB ID: 3R45).[16] In the Supporting Information, we present the same analysis of coarse-grained MD simulations based on the dimer subdomain of the octameric CENP-A nucleosome (PDB ID: 3AN2).[15] Currently, the CENP-A/H4/HJURP structure is the only one that includes the final six residues of CENP-A.

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

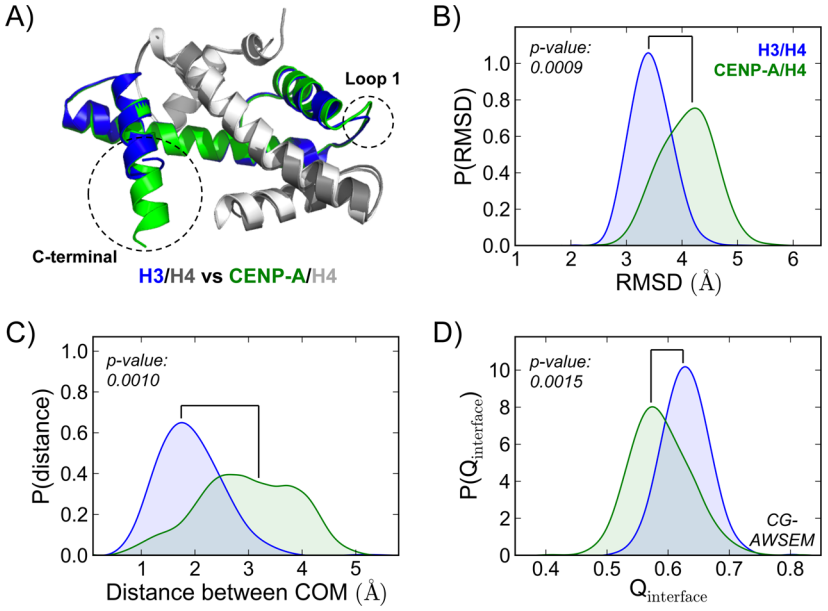


Figure 2. CENP-A/H4 displays greater structural variability than H3/H4 in CG-AWSEM simulations. (A) Structural alignment of CENP-A/H4 and H3/H4 highlights the two main structural differences between CENP-A and H3: the longer loop 1 and C-terminal regions of CENP-A (labeled by dashed circles). (B) Probability distribution functions of the Cα RMSD reveal that replacing H3 with CENP-A leads to greater structural variability in the dimer. (C) Probability distribution functions of the distance between the centers-of-mass (COM) of H3 (or CENP-A) and H4 show that CENP-A/H4 exhibits much more conformational heterogeneity. (D) Probability distribution functions of the $Q_{interface}$ with respect to the crystal structures of CENP-A/H4 (PDB ID: 3R45) and H3/H4 (PDB ID: 1AOI) for the CG-AWSEM simulation trajectories indicate that CENP-A/H4 has a more heterogeneous binding interface than H3/H4. Structure figure rendered in Pymol.

Distinguishing its structure from canonical H3, the C-terminal region of CENP-A is noted for its rapid evolution[12,44] and functionally required for binding to CENP-C.[45] Therefore, much of our analysis focuses on the C-terminal end of CENP-A.

Coarse-grained and all-atom results are presented separately in the following two sections. CG-AWSEM results characterize global features of CENP-A and H3 dimers, examining how the histone monomers contribute separately to dimer stability, comparing the structural variability of CENP-A/H4 and H3/H4, and investigating the effect of chaperone HJURP on the CENP-A/H4 dimer. Further, contacts analyses based on all-atom MD simulations in explicit solvent provide a detailed physical description of how HJURP interacts with the CENP-A dimer, mapping key contacts between HJURP and the C- and N-terminal portions of CENP-A.[46] Lastly, *in vivo* experiments investigate the role of CENP-A S68, testing the hypotheses derived from all-atom MD contact map analysis. We have found that both simulation methods reach the same overall consensus qualitatively when performing the same analyses. Global measures from all-atom simulations are presented in the Supporting Information.

CG-AWSEM MD Results. *H4 Adopts More Native-Like Conformations Than H3 and CENP-A.* All core histones share the "histone-fold" structural motif, three helices connected by two loops, yet the number of sequence variants for each differs widely. This difference has important implications for histone evolution[12] and nucleosome assembly dynamics. For instance, several variants exist for the canonical histone H3 (i.e., H3.1) including H3.2/H3.3/CENP-A,[6] while there are no variants for histone H4 reported thus far. From CG-AWSEM simulations, we first investigated how histone monomers H4 and H3, or H4 and CENP-A, contribute separately to dimer structural dynamics by calculating Q value, a normalized measure that compares the pairwise contacts in one structure to those in another (see Methods). A higher Q value (that can vary between 0 and 1) indicates greater structural similarity between the two structures. Here, we calculated the Q value between the simulation snapshots and the corresponding crystal structures for H3/H4 (PDB ID: 1AOI)[1] and CENP-A/H4 (PBD ID: 3R45).[16]

Interestingly, for all the systems studied, the conformations of H4 remain highly native-like, with an average Q value considerably greater than Q_{H3} or Q_{CENP-A}. The probability distributions of Q value for H4 are centered at ~0.8 (Figure 1A–D), corresponding to root-mean-squared deviations (RMSD) ranging from 1.7 to 2.1 Å, whereas Q value for H3 at 0.7 corresponds to a RMSD range from 2.0 to 2.6 Å and for CENP-A Q at 0.7 corresponds to a RMSD from 2.0 to 2.9 Å. H4 is consistently stable in both H3/H4 and CENP-A/H4 dimers, with and without the presence of chaperone HJURP; even though CENP-A displays large conformational variety in the CENP-A/H4 dimer, indicated by the broad distribution in $P(Q)$ (Figure 1C), H4 maintains native-like conformations for most of the simulation trajectories. When performing this analysis based instead on the CENP-A/H4 dimer found in the octameric CENP-A nucleosome crystal structure (PDB ID: 3AN2),[15] we reach the same conclusion (Figure S3). Histone H4 consistently maintains native-like stability, providing a strongly reinforcing structural framework for histone dimers and higher order structures, such as the histone octamer. The intrinsic stability of H4 is independent of its dimer partner, CENP-A or H3, or the presence of chaperone HJURP.

CENP-A/H4 Exhibits Greater Structural Variability than H3/H4. We then examined the structural variability of the CENP-A/H4 and canonical H3/H4 dimers in CG-AWSEM simulations by calculating the RMSD of Cα atoms with respect to the corresponding crystal structures. Replacing canonical H3

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207–13218

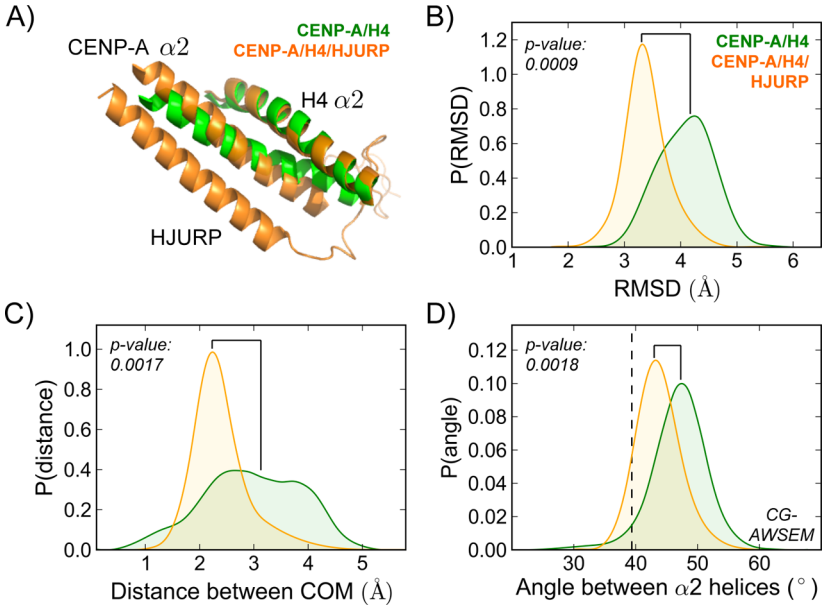


Figure 3. HJURP stabilizes the overall shape of the CENP-A/H4 dimer in CG-AWSEM simulations. (A) Representative simulation snapshots of CENP-A/H4 (green) and CENP-A/H4 in conjunction with HJURP (orange) illustrate how HJURP adjusts the overall shape of the dimer. Only the α2 helices of CENP-A and H4, as well as HJURP, are displayed. Introducing the CENP-A-specific chaperone HJURP (B) reduces the CENP-A/H4 RMSD, on average, with respect to the crystal structure and (C) reduces the average distance between the COMs of CENP-A and H3, focusing the distribution and making the CENP-A/H4 dimer more compact and stable. (D) HJURP modifies the overall shape of the CENP-A/H4 dimer by reducing the angle between the α2 helices of CENP-A and H4. The reference angle from the crystal structure (40°) is illustrated by the dashed line. Structure figures rendered in Pymol. Similar analyses for the all-atom simulations can be found in Figure S8.

with CENP-A in the heterodimer leads to a greater RMSD, on average, for both CG (Figure 2) and all-atom MD simulations (Figure S4). In the context of CG simulations, CENP-A/H4 (4.1 ± 0.5 Å) exhibits greater RMSD on average than H3/H4 (3.4 ± 0.4 Å) (Figure 2B). As expected, the two-residue longer loop 1 in CENP-A displays enhanced fluctuations (Figure S7).

The spontaneous variability of CENP-A/H4 dimer in CG simulations is not only due to its flexible loop 1. The distance between the centers-of-mass (COM) of CENP-A and H4 occupies a much broader distribution than H3 and H4 (Figure 2C), indicating that the interface between CENP-A and H4 is more globally flexible. We analyzed the binding interface by calculating $Q_{interface}$, a normalized measure comparing the interface contacts in the CG simulation snapshots to those in the crystal structures (PDB IDs 1AOI for H3/H4 and 3R45 for CENP-A/H4). As shown in Figure 2D, the distribution of the CENP-A dimer $Q_{interface}$ is shifted considerably to the left of the same distribution for the H3 dimer, demonstrating that substituting canonical H3 with CENP-A leads to less native-like interfaces and increases the conformational heterogeneity of the dimer binding interface. Additionally, we calculated the pairwise Q value between any two conformations within one simulation trajectory. As shown in Figure S6, the pairwise Q is greater on average for H3/H4 (0.81 ± 0.04) than for CENP-A/H4 (0.73 ± 0.08) in CG simulation, implying that the higher heterogeneity of CENP-A/H4 is intrinsic and spontaneous. Overall, the isolated CENP-A/H4 dimer is more structurally variable than H3/H4 in both CG-AWSEM and all-atom simulations. These data are consistent with the greater heterogeneity seen in the CENP-A nucleosome compared to its canonical H3 counterpart *in silico*, *in vitro*, and *in vivo*.[29,35,36]

HJURP Alters the Shape of the CENP-A/H4 Dimer. The data above demonstrate that, in isolation, the CENP-A/H4 dimer is structurally more variable than H3/H4 in CG simulations, which leads to the question of whether its chaperone HJURP influences the structural features of CENP-A/H4. Upon the introduction of HJURP, the RMSD distribution of the CENP-A dimer becomes tighter and shifts to the left (Figure 3B), centered at 3.3 Å, which is comparable to the RMSD of H3/H4 in isolation (Figure 2C). Moreover, the distance between CENP-A and H4 shows much less deviation when HJURP is present (Figure 3C). Therefore, in agreement with its documented role as a bonafide chaperone, HJURP stabilizes and restrains the conformational variability of the CENP-A/H4 dimer on a global scale.

Among the three major helices of each core histone, α2 is the longest helix and provides the main supportive frame for the histone-fold structure. Thus, the shape of the CENP-A/H4 dimer can be characterized on a coarse level by the angle between the α2 helices of CENP-A and H4. Introducing the CENP-A-specific chaperone HJURP reduces the average angle between the α2 helices of CENP-A and H4 by 6° (Figure 3D). The presence of HJURP tightens this distribution and brings its center closer to the reference value calculated from the crystal structure. As shown in the representative snapshot (Figure 3A), HJURP modifies the orientation of CENP-A with respect to H4, bringing the CENP-A dimer's structure closer to that found in its octameric nucleosome. When performing the same analysis for all-atom MD simulations, we observe that the introduction of HJURP slightly reduces the average RMSD (Figure S8A). However, the distance between histone monomers and the angle between α2 helices remain unchanged (Figure S8B,C). While CG-AWSEM MD simulations can explore conformational space widely, all-atom MD mainly probes dynamics near the native state, keeping global preferences relatively constant. Taken together, these results indicate that HJURP stabilizes the conformational ensemble of the CENP-A dimer and modifies the overall shape of CENP-A/

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

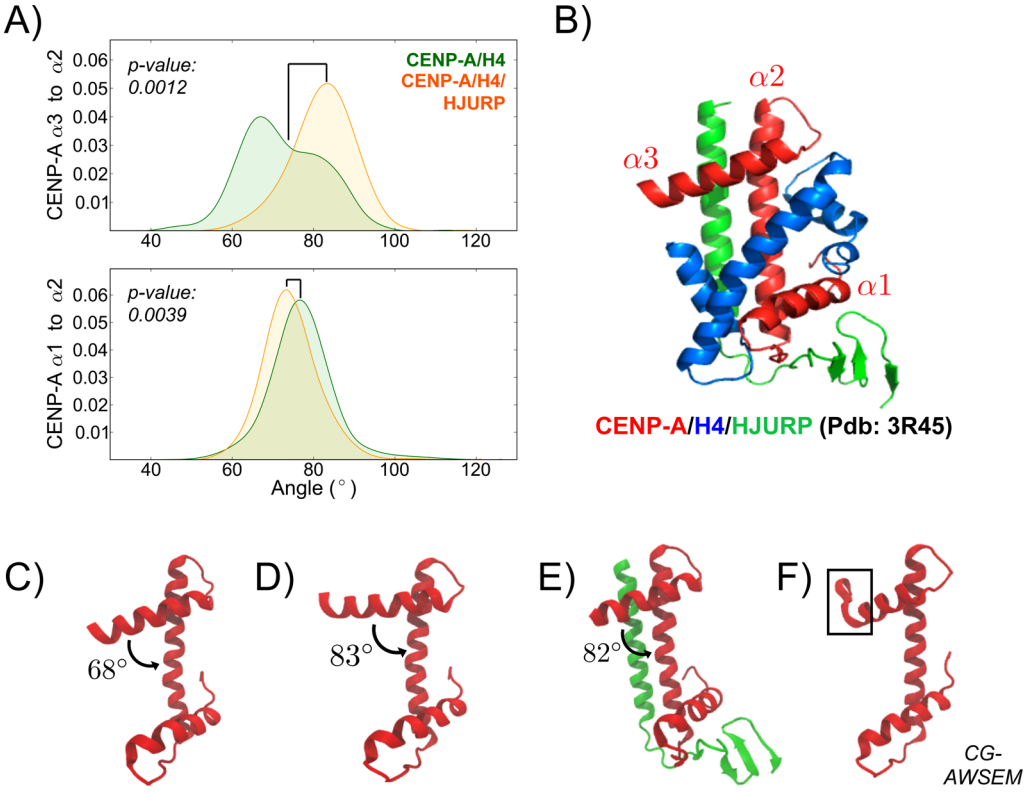


Figure 4. HJURP stabilizes CENP-A α3 in CG-AWSEM simulations. (A) Probability distributions of the angles between CENP-A α2 and α3 and between α1 and α2, demonstrate that the introduction of the chaperone HJURP stabilizes the motion of CENP-A α3 with respect to CENP-A α2. (B) The CENP-A/H4/HJURP crystal structure is shown. Helices used for the angle measurements are labeled in red. Conformations (C) and (D) correspond to the primary peak and shoulder in the distribution of the angle between α2 and α3 of CENP-A in the absence of HJURP. (E) A representative structure illustrates the most common angle between CENP-A α2 and α3 upon the introduction of HJURP. (F) In the absence of HJURP, the C-terminal end of α3 of CENP-A becomes partially unwound. Colors identify CENP-A (red) and HJURP (green). H4 is removed from the representative structures to facilitate easier observation. Structure figures rendered in VMD. Related CG trajectories can be found in the Supporting Information (Movies S1 and S2). We observe the same overall trend when analyzing the angles between α2 and α3 and between α1 and α2 of CENP-A in the all-atom MD simulations (Figure S9).

H4, priming the CENP-A/H4 dimer for its deposition into the nucleosome and, ultimately, into the centromere.

HJURP Regulates the CENP-A/H4 Dimer through Stabilizing the C-Terminal Helix of CENP-A. After investigating how the introduction of HJURP influences the CENP-A dimer structure globally, we turn our focus to how HJURP specifically modifies the conformational preferences of the CENP-A monomer. The CENP-A α3 helix includes the final six residues at the C-terminus (i.e., LEEGLG in the human CENP-A sequence, Figure S1), which are currently thought to play an important role in CENP-A's interaction with the chaperone HJURP[16] and kinetochore protein CENP-C.[45,47] Presently, only the CENP-A/H4/HJURP complex includes an ordered CENP-A C-terminus in its crystal structure. Therefore, to better understand how HJURP dynamically affects the α3 helix of CENP-A, we measured the angles between the CENP-A α1 and α2 helices and between CENP-A α3 and α2 (Figure 4B).

The α3−α2 angle of CENP-A is broadly distributed, with a primary peak and a shoulder, at ∼68° and ∼82° respectively (Figure 4A), corresponding to two populated states of CENP-A conformations when HJURP is absent (Figure 4C,D). However, in the presence of HJURP, this angular distribution becomes tightened exclusively around the 82° peak (Figure 4A,E). The preceding $Q_{monomer}$ analysis (Figure 1C,D) also illustrates the change of $Q_{CENP\text{-}A}$ from two populated states to one upon the introduction of HJURP. We observe the same overall trend in the all-atom MD results: The addition of HJURP stabilizes the angle between CENP-A α helices 2 and 3 (Figure S9A) without having a significant effect on the angle between CENP-A α1 and α2 (Figure S9B), in part because CENP-A α3 becomes partially unraveled in the absence of HJURP (Figure S9C).

The CENP-A α3 helix is much more structurally dynamic than α1 in the CG simulations, since the CENP-A α1−α2 angle occupies only one focused peak and remains unchanged upon the introduction of HJURP (Figure 4A). Further analysis reveals that the flexible CENP-A α3 helix could disrupt the stability of H4 α3 (Figure S11), which is consistent with all-atom contact maps (Figure 5). These results are also consistent with the experimentally determined B-factor data (Figure S10), which describes the uncertainty about the actual atom positions in X-ray crystallography. Moreover, these data provide a physical explanation of a key result from our previous CENP-A nucleosome work[35]—the shearing motion of the CENP-A nucleosome dimerization interface—wherein the interface, called the "four-helix bundle", is exactly defined by two copies of the CENP-A α3 and α2 helices. Altogether, our CG-AWSEM simulations demonstrate that HJURP regulates the CENP-A/H4 dimer through stabilizing the α3 helix of CENP-A.

All-Atom MD Results. *HJURP Facilitates the Formation of a Structure-Inducing Electrostatic Network with the C-*

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

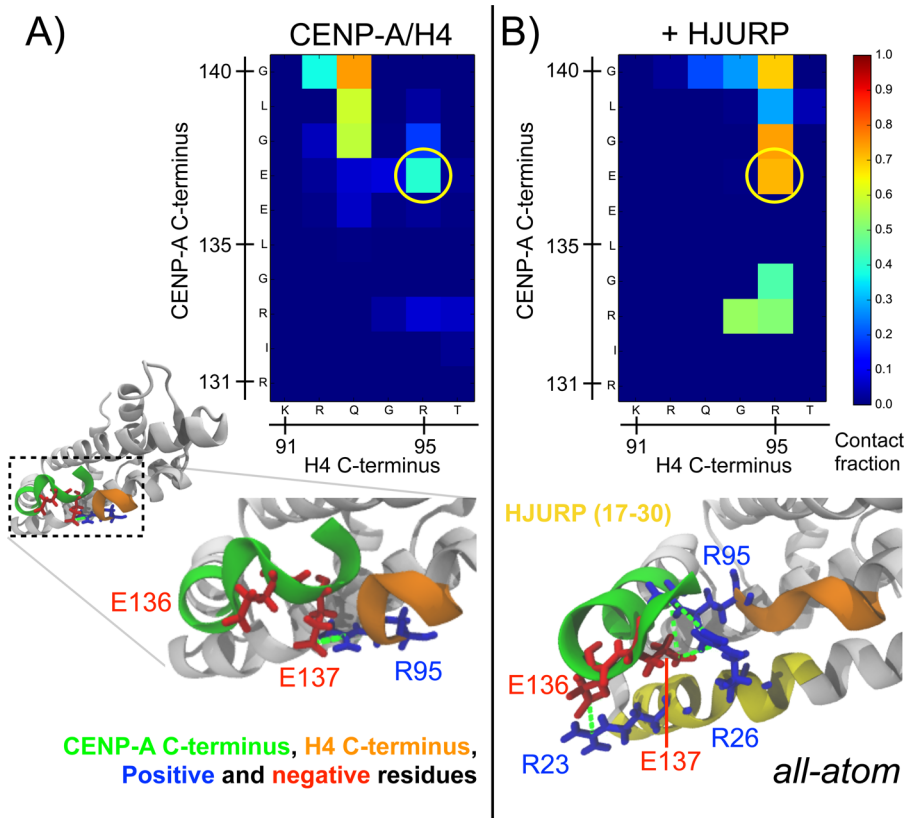


Figure 5. The presence of HJURP rearranges interactions between the C-termini of CENP-A and H4. Contact maps between the C-termini of CENP-A and H4, and representative simulation snapshots, in (A) the CENP-A/H4 dimer and in (B) the CENP-A/H4 dimer in conjunction with CENP-A specific chaperone HJURP illustrate that HJURP facilitates electrostatic interactions that introduce greater helical structure to the C-terminus of CENP-A. The solid yellow circle highlights a potentially critical salt-bridge between CENP-A and H4.

Termini of CENP-A and H4. After analyzing global conformational features in CG-AWSEM simulations, we examined finer details of the interactions between CENP-A and H4, and those between HJURP and CENP-A, in all-atom simulations. First, we mapped the contacts between the C-termini of CENP-A and H4 in the absence and presence of HJURP (Figure 5A,B). In the absence of HJURP, ∼40% of the time, a contact forms between the oppositely charged H4 R95 and CENP-A E137 (Figure 5A), and the α3 regions of CENP-A and H4 become partially unraveled. The C-terminal tail of CENP-A (the final 6 residues: 135−140) is ∼4% helical on average in the all-atom MD trajectory. The introduction of HJURP facilitates the formation of an electrostatic network between the C-termini of CENP-A and H4 and the α helix of HJURP, the contact between H4 R95 and CENP-A E137 increases to ∼70% (Figure 5B), and the α3 regions of CENP-A and H4 retain their helical structure. The C-terminal tail of CENP-A increases to ∼35% helical on average in the presence of HJURP. Therefore, HJURP regulates the electrostatic interactions and drives the helicity in the CENP-A C-terminus. These results are consistent with the crystallographic information; except for the CENP-A/H4/HJURP complex, all other CENP-A-included crystal structures published thus far do not include the final six residues of CENP-A, because these residues remain disordered in these structures.[15,25,48]

The C-terminal tail of CENP-A (-LEEGLG) carries an overall net charge of $-2e$ and is three residues longer than the corresponding neutral tail of H3 (-ERA). The increased acidity and length of the CENP-A C-terminal tail compared to H3 could play an important role in differentiating assembly chaperones and binding partners for these two histones. Indeed, as can be seen in the contact maps analysis, several charged residues, including HJURP R23, R26, CENP-A E136, E137, and H4 R95, form a network of interactions at the interface between the C-terminus of CENP-A, the C-terminus of H4, and the α helix of HJURP (Figure 6B). In contrast, H3/H4 does not form analogous interactions upon the introduction of HJURP (Figure 6A). Thus, the neighboring acidic residues near the C-terminus of CENP-A (E136 and E137) allow CENP-A to form key electrostatic interactions with basic residues of H4 (R95) and HJURP (R23 and R26).

CENP-A Forms Key Interactions with the Hydrophobic β Domain of HJURP. On the other side, the N-terminal portion of the CENP-A histone-fold interacts with the hydrophobic β domain of HJURP. Previous experimental studies have focused on the role of CENP-A S68 in HJURP recognition, which has been challenged.[16,49,50] Here, we performed contact map analysis of the CENP-A/H4/HJURP all-atom simulations to examine the contribution of CENP-A S68 in atomistic detail. These analyses reveal that CENP-A S68 inserts well into the hydrophobic pocket formed by the β domain of HJURP (V50, M52, L55, and W66) (Figure 7B). On the contrary, H3 Q68 almost exclusively interacts with HJURP W66, leading to a closed hydrophobic pocket (Figure 7A). While CENP-A S68 and L91 both form contacts with the hydrophobic pocket, there are virtually no interactions between these two CENP-A residues (only ∼2%). However, H3 Q68 interacts significantly with H3 V89 (∼20%), which is the H3 analogue of CENP-A L91. The data suggest that the shorter side chain of CENP-A S68 cannot reach CENP-A L91, whereas H3 Q68 is long

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

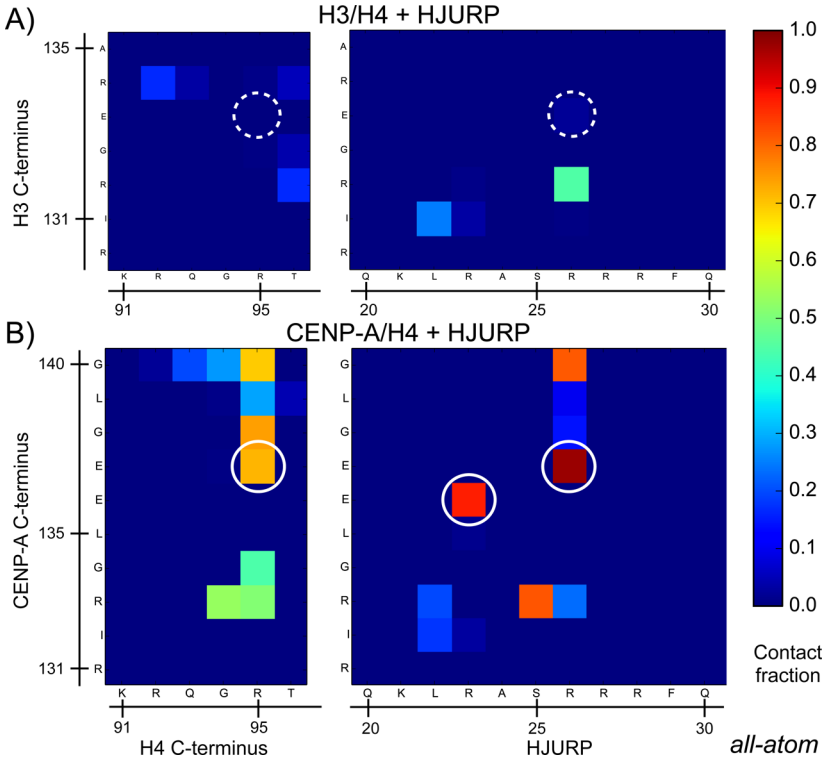


Figure 6. HJURP forms electrostatic interactions with the C-termini of CENP-A/H4, but not H3/H4. (A) The H3 C-terminus does not form significant interactions with the H4 C-terminus and α helix of HJURP in the H3/H4/HJURP all-atom trajectory. (B) Contact maps of the C-terminal region of CENP-A with the C-terminus of H4 and the α helix of HJURP in the all-atom simulation of CENP-A/H4/HJURP identify key electrostatic interactions. Solid white circles highlight specific salt-bridges, and dashed circles represent the lack thereof.

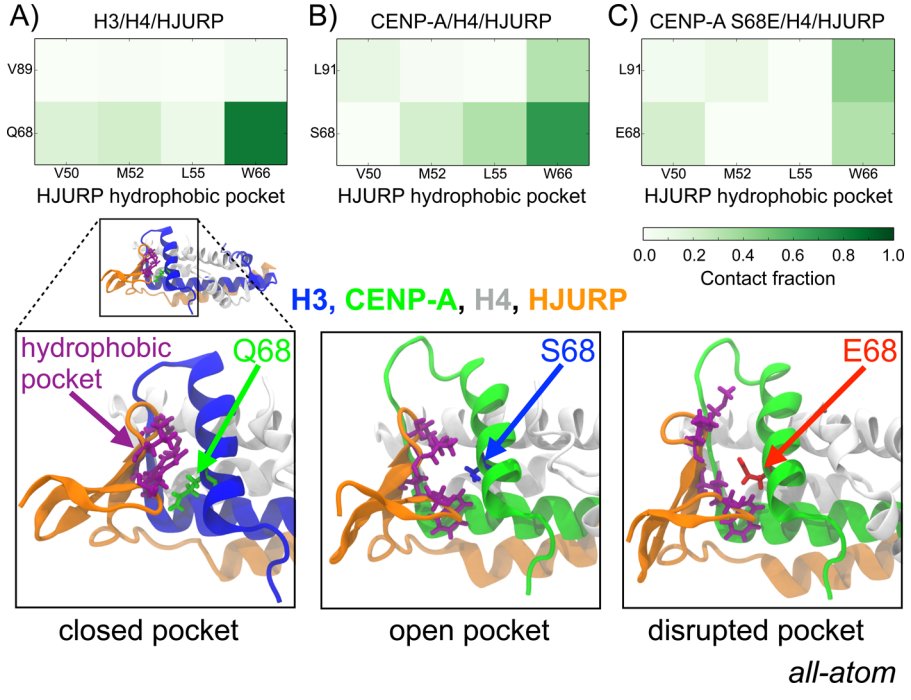


Figure 7. CENP-A forms key interactions with the hydrophobic pocket of HJURP. Contact maps between the hydrophobic pocket of HJURP (i.e., V50, M52, L55, and W66; in purple tubes) and key residues of (A) canonical H3, (B) CENP-A, and (C) CENP-A, where S68 is replaced with E68 display different types of interactions. H3 Q68 almost exclusively interacts with HJURP W66, and HJURP's pocket becomes closed. CENP-A S68 forms contacts with multiple residues of the hydrophobic pocket, which remains open. When replacing CENP-A S68, E68 (shown in red tubes) disrupts the interactions between CENP-A and the hydrophobic pocket of HJURP. Colors identify H3 (blue), CENP-A (green), and HJURP (orange). Structure figures rendered in VMD.

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

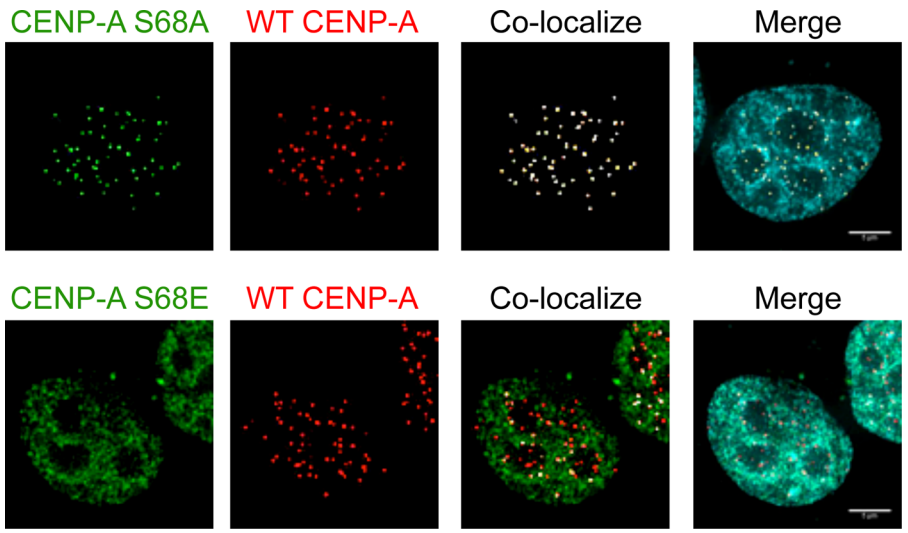


Figure 8. CENP-A S68A localizes to the centromere, whereas CENP-A S68E does not. Residue S68 in CENP-A is mutated to alanine or glutamic acid, respectively. Mutants are GFP-tagged and co-expressed with mCh-tagged WT CENP-A to assess co-localization. Co-localized foci appear as white dots in the co-localized column. Merge column shows the DAPI-stained DNA within the nucleus.

enough to form contacts with H3 V89. Furthermore, since H3 Q68 and H3 V89 interact with each other, they cannot both insert simultaneously into the HJURP hydrophobic pocket (Figure 7A). Between CENP-A S68 and CENP-A L91, S68 is more dominant in binding to HJURP: CENP-A S68 forms a contact with HJURP W66 85% of the time, while the contact between CENP-A L91 and HJURP W66 is only present ∼35% of the time (Figure 7B). Together, due to side chain lengths and strong to moderate hydrophobicities, CENP-A S68 and L91 permit CENP-A to form stronger interactions with HJURP than H3 Q68 alone.

To test our hypothesis that CENP-A S68 is required to bind with HJURP due to both the short length and some hydrophobicity (and electric neutrality) of its side chain, we performed *in vivo* experiments and all-atom simulations mutating this residue. Alanine (A), which is short and hydrophobic, and glutamic acid (E), which is long and negatively charged, served as valuable replacement residues, denoted CENP-A S68A and S68E, respectively. In the experiment, we aimed to determine whether the S68-mutated CENP-A could still be functionally deposited to the centromeric region by its chaperone HJURP *in vivo*. Successful binding with HJURP drives CENP-A deposition exclusively to the centromeres, whereas disrupted binding with HJURP is predicted to lead to the ectopic deposition of CENP-A. Site-directed mutagenesis experiments were conducted for CENP-A S68A and CENP-A S68E. These GFP-tagged CENP-A S68 mutants were co-expressed with mCh-tagged wild-type (WT) CENP-A under the control of a constitutive promoter, and the mutants' ability to localize to either the centromere or at the ectopic regions was determined. Comparing the localization of mutated and WT CENP-A (Figure 8), it can be seen that the mutant CENP-A S68A results in robust centromeric localization, while the mutant CENP-A S68E is not localized to the centromeres but displays ectopic incorporation.

To gain more biochemical insight into the specific role of S68, we performed all-atom simulations of CENP-A/H4/HJURP replacing CENP-A serine 68 with glutamic acid. The CENP-A S68E mutant disrupts the interactions between CENP-A and the hydrophobic pocket of HJURP (Figure 7C). The longer side chain of E68 sterically clashes with HJURP's hydrophobic pocket, pushing it away from the CENP-A α1 helix. Once pushed away, the hydrophobic pocket becomes disrupted and loses its structural integrity. This explains why S68E CENP-A cannot successfully be recognized and loaded by chaperone HJURP in our *in vivo* experiments. Overall, our all-atom MD simulations and *in vivo* experiments demonstrate that CENP-A S68 is necessary to maintain the unique binding interface between CENP-A and the hydrophobic β domain of HJURP. All-atom simulation results indicate that the short length of S68's side chain is essential for CENP-A's recognition by the hydrophobic β domain of HJURP.

DISCUSSION

In this report, coarse-grained and all-atom MD simulations provide a dual-resolution perspective of the effects of HJURP and CENP-A on histone dimer dynamics. These data reveal that the replacement of canonical H3 with CENP-A translates into increased conformational heterogeneity in histone dimer dynamics (Figure 2). Furthermore, the chaperone HJURP plays a stabilizing role for the CENP-A/H4 dimer and modifies the CENP-A dimer's overall shape (Figure 3) as a potentially priming step in advance of the CENP-A loading. H4 remains stable and adopts native-like conformations in both CENP-A/H4 and H3/H4 (Figure 1). This intriguing distinction is consistent with the fact that H4 remains conserved throughout eukaryotic evolution, whereas distinct variants of H3 exist for special roles in transcription and chromosome segregation. Thus, H4 could provide a consistent reinforcing structural framework for histone dimers, while the H3 family, including canonical H3 and the centromere-specific variant CENP-A, provides variability to the structure and function.

Our overarching aim is to investigate the fundamental dynamics of the histone dimers H3/H4 and CENP-A/H4. Therefore, only the histone-fold domains were previously considered, excluding the H3 (CENP-A) N-terminal helix and histone tails, based on the fact that those regions are primarily involved in the interactions with DNA or other histones, such as H2A/H2B (Figure S14). Nevertheless, in the nucleosome

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

structure, the H4 C-terminal tail forms a few hydrophobic interactions with H3 (CENP-A) α2 and H4 α3, suggesting the possibility that the H4 C-terminal tail stabilizes histone dimers (Figure S14). In CG simulations, the angle between CENP-A α2 and H4 α3 is mostly stable in the absence of the H4 C-terminal tail (Figure S11). Further CG simulations demonstrate that including the H4 C-terminal tail increases the structural flexibility of the CENP-A/H4 dimer, compared to when the H4 C-tail is excluded (Figures S3B,D and S15B,D). It is feasible that H2A/H2B, together with H3(CENP-A) α2 and H4 α3, stabilizes the H4 C-terminal tail, as can be seen in the nucleosome crystal structure: β strands form between the H4 C-terminal tail (H4 T96 and Y98) and H2A T101 (Figure S14). Interestingly, even with the H4 C-terminal tail included, H4 still adopts more native-like conformations than CENP-A (Figures S3C and S15C). Investigating the precise role of histone tails in the CENP-A/H4/HJURP complex and the structural dynamics comparison between CENP-A/H4 and H3/H4 homotypic or heterotypic histone tetramers are important future directions.

The variability of CENP-A is due, in part, to its longer C-terminal residues (six in CENP-A versus three in H3), which maintain helical structural integrity only when in a complex with HJURP (Figure 5). The increased acidity of the CENP-A C-terminus ($-2e$) compared to the neutral charge of the corresponding C-terminus in H3 could contribute to HJURP's specificity to CENP-A.[45] The coarse-grained MD results demonstrate that HJURP reduces the conformational heterogeneity of the CENP-A/H4 dimer by modifying the dimer's overall shape and stabilizing the CENP-A α3 helix (Figures 3 and 4). Furthermore, all-atom MD simulations illustrate that HJURP forms a structure-inducing electrostatic network with the C-termini of CENP-A and H4 but not with H3/H4 (Figures 5 and 6). The two-residue-longer loop 1 region of CENP-A is subject to less fluctuations upon the introduction of HJURP (Figure S7), which indicates that HJURP stabilizes loop 1 region of CENP-A indirectly. Debate continues over the role of CENP-A S68[16,49,51] and its post-translational modification[46] in CENP-A's interaction with HJURP and deposition into the nucleosome. Replacing CENP-A S68 with E68 *in vivo* and in all-atom MD simulations mimics S68 phosphorylation by elongating the side chain and introducing a negative charge. Recent studies suggest that phosphorylating S68 is sufficient to disrupt CENP-A–HJURP binding. In our experiments (Figure 8), mutating this residue to glutamic acid resulted in ectopic CENP-A deposition *in vivo*. All-atom simulations provide a physical explanation of how S68 phosphorylation could disrupt the binding interface between CENP-A and HJURP: when replacing CENP-A S68, the longer E68 side chain sterically clashes with HJURP's hydrophobic pocket, pushing it away from the CENP-A α1 helix and disrupting the pocket's overall shape. Together, *in vivo* and all-atom simulation results support the previously proposed model in which CENP-A S68 phosphorylation (S68ph) must be tightly regulated, and the eviction of CENP-A's chaperone HJURP must be orchestrated within a small window of the cell cycle in order to minimize the risk of ectopic CENP-A incorporation.[46]

Further analysis reveals that the introduction of HJURP to H3/H4 significantly disrupts the binding interface between H3 and H4 (Figure S12B) and leads to a slightly larger average RMSD in CG-AWSEM simulations (Figure S12A), compared to the H3/H4 dimer in isolation. In all-atom simulations of the same system, the introduction of HJURP destabilizes a key electrostatic interaction between the C-termini of H3 and H4 (Figure S13). These results may provide a partial explanation for experimental evidence suggesting that H3/H4 cannot bind HJURP *in vitro*.[38,41,49]

Based on our observations above, it is possible that a currently under-appreciated role for chaperone HJURP may also be its ability to "lock" the C-terminus of CENP-A before it encounters another kinetochore protein. HJURP may work as a switch, turning on and off the binding availability of the CENP-A C-terminal tail. The presence of HJURP stabilizes the C-terminus of CENP-A before CENP-A's deposition, and after CENP-A is deposited, HJURP must release the intrinsically disordered C-terminal tail of CENP-A, in order for it to become available to bind with another kinetochore protein, most critically, CENP-C.[36,45] The structural alignment of CENP-A from different molecular contexts clearly shows the "on" and "off" states of its C-terminal tail (Figure S16). Plus, recent research by Tachiwana et al. illustrates that CENP-C recruitment requires direct interaction between CENP-C and HJURP.[52] Consequently, HJURP may be unique in that it functions as a protein-folding chaperone for CENP-A, stabilizing the CENP-A/H4 dimer, and also as a protein-binding chaperone for CENP-C and CENP-A, mediating CENP-C's recruitment to the CENP-A nucleosome. A related work previously reported on the interaction between the chaperone Chz1 and the H2A.Z/H2B dimer, wherein the chaperone Chz1 undergoes a disorder-to-order transition upon binding to H2A.Z/H2B,[9] suggesting such transitions might be conserved in the structure-inducing mechanisms employed by histone chaperones.[53−55]

The dual-resolution nature of this study provides a unique opportunity to directly compare and cross-validate the same results from both CG and all-atom simulations. Therefore, for each of the main CG results (monomer flexibility; dimer variability; global shape; and HJURP's effect on the angle between helices), we performed the same analysis on the all-atom MD trajectories, including the resulting figures in Figures S2, S4, S8, and S9. Overall, all-atom and CG methods reach the same consensus qualitatively. However, how the results of these two techniques differ is important to our work as well. When examining global properties including pairwise Q, interface Q, and the distances between histones, the results based on all-atom MD simulations remain close to the native state, and these properties do not vary much across different systems. On the other hand, the analysis of CG simulations reveals significant differences in the global properties of the systems studied, clearly illustrating the value added by including CG simulations. The strength of all-atom MD lies in its ability to probe specific interactions and native-state dynamics at high resolution. For example, when replacing CENP-A S68 with E68 in all-atom simulations, the glutamic acid sterically clashes with HJURP's hydrophobic pocket, pushing the pocket away from the CENP-A α1 helix (Figure 7). This detailed effect is not observed in CG-AWSEM MD simulations because it is mainly due to the long length of the glutamic acid side chain, a difficult property to capture in a three-bead per amino acid model. Altogether, CG explores greater conformational space at a more global level, and all-atom MD investigates finer details close to the native state.

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

CONCLUSION

Our dual-resolution MD simulations shed light on the differences between the structural dynamics of the CENP-A/H4 and H3/H4 dimers, providing insight into how HJURP primes the CENP-A/H4 dimer for deposition. Our results indicate that HJURP, while potentially acting as a disruptive force for H3/H4, serves as a protein-folding chaperone for the CENP-A dimer and a protein-binding chaperone for CENP-C and the CENP-A dimer. Finally, this study makes predictions about the key histone−histone and CENP-A-HJURP interactions, one of which is confirmed by *in vivo* experiments and provides new dynamic insights into the underlying mechanisms governing the HJURP-mediated assembly of CENP-A nucleosomes *in vivo*.

METHODS

Structure Preparation for MD Simulations. Starting from the crystal structures for canonical H3 nucleosome (PDB ID: 1AOI)[1] and the CENP-A/H4 heterodimer with chaperone HJURP (PDB ID: 3R45),[16] we developed all-atom and CG-AWSEM models for four systems: (1) the H3/H4 heterodimer; (2) the CENP-A/H4 heterodimer; (3) the H3/H4 heterodimer with the CENP-A specific chaperone HJURP (as a control); and (4) the CENP-A/H4 heterodimer in a complex with the chaperone HJURP. Systems 1, 2, and 4 are based directly on PDB structures, or subdomains thereof, and we aligned the H3/H4 dimer to the CENP-A/H4 dimer of CENP-A/H4/HJURP to construct a CG-AWSEM model for H3/H4 in conjunction with HJURP. Finally, for the all-atom model of H3/H4/HJURP, we rotated the final three residues of H4 (-GRT) slightly after alignment to the CENP-A dimer in order to prevent structural overlaps between H4 and the newly placed HJURP. From these four models, at two different resolutions, we performed all-atom and coarse-grained MD simulations.

The CENP-A/H4/HJURP crystal (PDB: 3R45) does not include the H4 C-terminal tail, but in the nucleosome structure, the H4 C-terminal tail is resolved and forms a few hydrophobic interactions with H3 (CENP-A) α2 and H4 α3 (Figure S14). Additional CG simulations were performed for a mixed CENP-A/H4, where CENP-A is provided from CENP-A/H4/HJURP (PDB: 3R45) and H4 from the CENP-A nucleosome (PDB: 3AN2), and for a CENP-A/H4 dimer derived solely from the CENP-A nucleosome structure (Figures S3 and S15). Both simulations demonstrate that the H4 C-terminal tail is intrinsically unstable. The results of these additional runs are addressed in the Discussion section and presented in the Supporting Information.

All-Atom MD Methods. We performed all-atom MD in explicit solvent using the gromacs 4.5.7 MD software,[56] the amber99SB*-ILDN[57,58] force field for proteins, the ions08[59] force field for ions, and the TIP3P water model. Using the *pdb2gmx* tool in Gromacs, we set the Lys and Arg residues to +1*e*, the Asp and Glu residues to −1*e*, the Gln residues to neutral, and protonated the His residues solely at NE2. Each system was solvated in a cubic water box, ensuring a minimum buffer length of 15 Å between the system and the edges of the box. We introduced Na^+ and Cl^- ions to neutralize the charge and represent the physiological 0.150 M NaCl environment. The systems were minimized using steepest descent, until reaching a maximum force <100 kJ/(mol nm). Periodic boundary conditions were employed throughout all the simulations, and long-range electrostatics were treated with the particle mesh Ewald method.[60] Nonbonded Coulomb and Lennard-Jones interactions were truncated at 10 Å, and all bonds involving hydrogen were constrained using the LINCS[61] algorithm. After minimization, the systems were heated to 300 K by 500 ps of protein-restrained NVT MD simulation followed by 500 ps of NVT MD simulation without restraints. After reaching thermal equilibrium, the systems were equilibrated at 300 K and 1.0 bar for 1.5 ns in the NPT ensemble.

To characterize the structure and dynamics of the canonical and CENP-A heterodimers with and without the chaperone HJURP, we performed unrestrained production all-atom MD simulations in the NPT ensemble at 1.0 bar and 300 K with a 2 fs time-step, saving coordinates, velocities, and energies every 2 ps for further analysis. We updated the list of nonbonded neighbors every 10 steps. For each system, 1 μs of MD simulations was performed using the V-rescaled, modified Berendsen thermostat[62] with a 0.1 ps time-constant and the Parrinello−Rahman barostat[63] with a relaxation time of 2.0 ps. For analysis, we only considered the final 600 ns of the trajectories to account for further temperature and pressure equilibration. Convergence of the all-atom simulations can be seen from the RMSD (Figure S5) and root-mean-square-inner-product (RMSIP)[64,65] analysis (Figure S17). A detailed explanation of the RMSIP calculation is provided in the Supporting Information.

Coarse-Grained MD Methods. For coarse-grained MD, we used associative memory, water-mediated, structure and energy model (AWSEM)[43] as the force field. In AWSEM, three beads, C_α, C_β (H for glycine), and O, represent one amino acid. Water-mediated interactions[66] are applied instead of other explicit or implicit water models. Fragment memory, which is included in the associate memory potential, is set as a single memory determined by the crystal structure of the corresponding histone monomer. Fragments are nonoverlapping and 12 (or fewer) residues long to ensure that it only provides a local structural bias. The interface dynamics between two molecules is purely determined by physics, not including any bioinformatics terms. To prevent the division of one dimer into two monomers, we applied a weak harmonic spring between the centers-of-mass of the two monomers (k = 0.02 kcal/(mol Å^2)). More details about AWSEM are included in the original force field study.[43]

AWSEM coarse-grained MD simulations are run through the LAMMPS package. Using the Nose−Hoover thermostat, we perform 200 ns NVT MD runs at 300 K with the initial velocities randomly generated for every bead drawn from a Maxwell−Boltzmann distribution. Five independent simulations with different random seeds of velocity distributions are carried out for each system. For analysis, we combine all five independent simulations after reaching equilibrated states, by deleting the first 10 ns, which is considered as the time required to reach equilibration (Figure S5). The trajectory is saved every 1000 time steps, which is 2 ps in the coarse-grained time scale. It is worth noting that the time scale in coarse-grained simulation is different from the time scale in all-atom simulation. Due to the faster diffusion, the same amount of CG-AWSEM simulation time samples much more conformational phase space than all-atom simulation does. CG simulations reach the convergence at around 10 ns, as shown in the RMSD and RMSIP analysis (Figures S5 and S17). It is important to note that while the time scale of atomistic simulations is absolute and can be directly related to experimental time scales, 10 ns of CG simulations cover several orders of magnitude longer real time scale (microsecond-to-millisecond).

***In Vivo* Experiments: Cloning and Immunofluorescence.** Original GFP-CENP-A and mCh-CENP-A plasmids were a gift from Stephan Diekmann. To generate the mutant serine 68, we performed fusion PCR with mutant forward primers ATAAGGAAGCTGCCCTTC[GCA]CGC or ATAAGGAAGCTGCCCTTC[GAA]CGC with a common reverse primer GAAGGGCAGCTTCCTTATCA for the [alanine] or [glutamic acid], respectively. The whole mutant CENP-A coding sequence after fusion PCR was cloned in-frame and downstream of the EGFP and linker peptide. The plasmids were cotransfected using Roche's X-tremeGENE HP DNA transfection reagent (cat. no. 06-366-546-001, lot no. 11062300) into HeLa cells that were grown on poly-D-lysine coated coverslips. Three days after transfection, the coverslips were cytospun at 800 rpm for 5 min to reduce the number of Z-stacks during immunofluorescence. Coverslips were then prefixed with 4% paraformaldehyde (PFA) for 1 min, washed 3× with PEM (80 mM K-PIPES, pH: 6.8; 5 mM EGTA, pH: 7.0; 2 mM $MgCl_2$), soluble proteins extracted with 0.5% Triton-X100 in CSK buffer (10 mM PIPES, pH: 6.8; 100 mM NaCl; 200 mM sucrose; 3 mM $MgCl_2$; 1 mM EGTA) for 5 min at 4 °C, washed once with PEM and fixed with 1% PFA for 20 min at 4 °C. The coverslip was then washed 3× with PEM, air-dried in the dark, and mounted with Vectashield with DAPI (softset) and sealed along the edges with

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

nail polish. Slides were stored in the dark at 4 °C until imaging with a DeltaVision RT system fitted with a CoolSnap charge-coupled device camera and mounted on an Olympus IX70.

Analysis for the MD Simulation Trajectories. We first determined the RMSD of all the Cα atoms of the CENP-A/H4 and H3/H4 dimers with respect to their corresponding crystal structures, investigating overall structural variation. We analyzed inter-residue contact preferences at the interface of CENP-A and H4, in the absence and presence of HJURP. A contact was determined to exist when the distance between two non-hydrogen atoms from different residues was <3.6 Å. Contacts were calculated as fractions of time of their respective entire trajectories. We used the STRIDE[67] algorithm to assign secondary structure to the all-atom simulation snapshots, considering the final six residues of CENP-A assigned as either 3_{10} or α to be helical. The average helical percentage was determined for each residue, and the average helicity of the CENP-A C-terminal tail was calculated as the mean of the averages for the final six residues.

To analyze the data from a more global perspective, we calculated a specific measure of structural similarity, Q,[68] of all the simulation snapshots to the experimentally determined crystal structures. A widely used quantity in protein folding theory, Q is a normalized order parameter, with higher values indicating greater structural resemblance between the two structures being compared:

$$Q = \frac{1}{n} \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{ij}^{\text{native}})^2}{2\sigma_{ij}^2}\right] \quad (1)$$

where n is the total number of contacts, r_{ij} is the instantaneous distance between the Cα atoms of residues i and j, r_{ij}^{native} is the same distance in the native state obtained from experiment, and σ_{ij} is a resolution parameter where $\sigma_{ij} = (1 + |i - j|)^{0.15}$. We generated probability density functions $P(Q)$ of all the simulation snapshots, where the shape of this distribution characterizes the structural heterogeneity of the related conformational ensemble. We first applied this order parameter to interface profiles of H3/H4 and CENP-A/H4. A pair of residues from CENP-A or H3 and H4 was considered a native contact if their Cα atoms are within 12 Å in the experimentally determined X-ray crystal structure, and only native interface contacts are considered for $Q_{\text{interface}}$ calculation. Lastly, we applied this formula of structural similarity to the native state to CENP-A or H3 and H4 histones separately, which we refer to as Q_{monomer}.

The angle between two α helices was determined by calculating the orientation vectors for selected helices. The assessment of convergence was mainly through RMSD and RMSIP. RMSIP was calculated using the first 10 eigenvectors of a given subspace. Detailed explanations of the methods used to determine helix orientation vectors and to calculate RMSIP values are provided in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.6b05355.

Further details on the conformational dynamics of histone dimers observed through all-atom and CG-AWSEM MD simulations (PDF)
Movie 1 (MPG)
Movie 2 (MPG)

■ AUTHOR INFORMATION

Corresponding Authors

*dalaly@mail.nih.gov
*gpapoian@umd.edu

Author Contributions

‖These authors contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

H.Z. is supported by the joint NCI-UMD Cancer Technology Partnership Program. D.W. is supported by the University of Maryland. M.B. and Y.D. are supported by the intramural research program of the CCR/NCI. G.A.P. is supported by the National Science Foundation NSF CHE-1363081. Computational resources are provided by the Deepthought-II supercomputer at the University of Maryland.

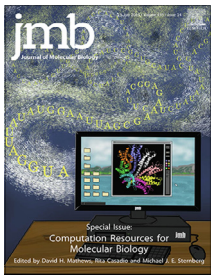
■ REFERENCES

(1) Luger, K.; Mader, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. *Nature* **1997**, *389*, 251−260.
(2) Arents, G.; Moudrianakis, E. N. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 11170−11174.
(3) Henikoff, S.; Furuyama, T.; Ahmad, K. *Trends Genet.* **2004**, *20*, 320−326.
(4) Sarma, K.; Reinberg, D. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 139−149.
(5) Melters, D. P.; Nye, J.; Zhao, H.; Dalal, Y. *Genes* **2015**, *6*, 751−776.
(6) Volle, C.; Dalal, Y. *Curr. Opin. Genet. Dev.* **2014**, *25*, 8−14.
(7) Biterge, B.; Schneider, R. *Cell Tissue Res.* **2014**, *356*, 457−466.
(8) Dalal, Y.; Furuyama, T.; Vermaak, D.; Henikoff, S. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 15974−15981.
(9) Chu, X.; Wang, Y.; Gan, L.; Bai, Y.; Han, W.; Wang, E.; Wang, J. *PLoS Comput. Biol.* **2012**, *8*, e1002608.
(10) Kamakaka, R. T.; Biggins, S. *Genes Dev.* **2005**, *19*, 295−316.
(11) Talbert, P. B.; Henikoff, S. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 264−275.
(12) Malik, H. S.; Henikoff, S. *Nat. Struct. Mol. Biol.* **2003**, *10*, 882−891.
(13) Tomonaga, T.; Matsushita, K.; Yamaguchi, S.; Oohashi, T.; Shimada, H.; Ochiai, T.; Yoda, K.; Nomura, F. *Cancer Res.* **2003**, *63*, 3511−3516.
(14) Lacoste, N.; Woolfe, A.; Tachiwana, H.; Garea, A. V.; Barth, T.; Cantaloube, S.; Kurumizaka, H.; Imhof, A.; Almouzni, G. *Mol. Cell* **2014**, *53*, 631−644.
(15) Tachiwana, H.; Kagawa, W.; Shiga, T.; Osakabe, A.; Miya, Y.; Saito, K.; Hayashi-Takanaka, Y.; Oda, T.; Sato, M.; Park, S.-Y.; Kimura, H.; Kurumizaka, H. *Nature* **2011**, *476*, 232−235.
(16) Hu, H.; Liu, Y.; Wang, M.; Fang, J.; Huang, H.; Yang, N.; Li, Y.; Wang, J.; Yao, X.; Shi, Y.; Li, G.; Xu, R.-M. *Genes Dev.* **2011**, *25*, 901−906.
(17) Shelby, R. D.; Vafa, O.; Sullivan, K. F. *J. Cell Biol.* **1997**, *136*, 501−513.
(18) Yoda, K.; Ando, S.; Morishita, S.; Houmura, K.; Hashimoto, K.; Takeyasu, K.; Okazaki, T. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 7266−71.
(19) Tanaka, Y.; Tachiwana, H.; Yoda, K.; Masumoto, H.; Okazaki, T.; Kurumizaka, H.; Yokoyama, S. *J. Biol. Chem.* **2005**, *280*, 41609−18.
(20) Camahort, R.; Shivaraju, M.; Mattingly, M.; Li, B.; Nakanishi, S.; Zhu, D.; Shilatifard, A.; Workman, J. L.; Gerton, J. L. *Mol. Cell* **2009**, *35*, 794−805.
(21) Dalal, Y.; Wang, H.; Lindsay, S.; Henikoff, S. *PLoS Biol.* **2007**, *5*, e218.
(22) Mizuguchi, G.; Xiao, H.; Wisniewski, J.; Smith, M. M.; Wu, C. *Cell* **2007**, *129*, 1153−64.
(23) Williams, J. S.; Hayashi, T.; Yanagida, M.; Russell, P. *Mol. Cell* **2009**, *33*, 287−98.
(24) Furuyama, T.; Henikoff, S. *Cell* **2009**, *138*, 104−113.
(25) Sekulic, N.; Bassett, E. A.; Rogers, D. J.; Black, B. E. *Nature* **2010**, *467*, 347−351.
(26) Dechassa, M. L.; Wyns, K.; Li, M.; Hall, M. A.; Wang, M. D.; Luger, K. *Nat. Commun.* **2011**, *2*, 313.
(27) Zhang, W.; Colmenares, S. U.; Karpen, G. H. *Mol. Cell* **2012**, *45*, 263−269.
(28) Shivaraju, M.; Unruh, J. R.; Slaughter, B. D.; Mattingly, M.; Berman, J.; Gerton, J. L. *Cell* **2012**, *150*, 304−316.

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218

(29) Bui, M.; Dimitriadis, E. K.; Hoischen, C.; An, E.; Quénet, D.; Giebe, S.; Nita-Lazar, A.; Diekmann, S.; Dalal, Y. *Cell* **2012**, *150*, 317−326.
(30) Furuyama, T.; Codomo, C. A.; Henikoff, S. *Nucleic Acids Res.* **2013**, *41*, 5769−5783.
(31) Hasson, D.; Panchenko, T.; Salimian, K. J.; Salman, M. U.; Sekulic, N.; Alonso, A.; Warburton, P. E.; Black, B. E. *Nat. Struct. Mol. Biol.* **2013**, *20*, 687−695.
(32) Wisniewski, J.; Hajj, B.; Chen, J.; Mizuguchi, G.; Xiao, H.; Wei, D.; Dahan, M.; Wu, C. *eLife* **2014**, *3*, e02203.
(33) Henikoff, S.; Ramachandran, S.; Krassovsky, K.; Bryson, T. D.; Codomo, C. A.; Brogaard, K.; Widom, J.; Wang, J.-P.; Henikoff, J. G. *eLife* **2014**, *3*, e01861.
(34) Black, B. E.; Foltz, D. R.; Chakravarthy, S.; Luger, K. *Nature* **2004**, *430*, 578−582.
(35) Winogradoff, D.; Zhao, H.; Dalal, Y.; Papoian, G. A. *Sci. Rep.* **2015**, *5*, 17038.
(36) Falk, S. J.; Guo, L. Y.; Sekulic, N.; Smoak, E. M.; Mani, T.; Logsdon, G. A.; Gupta, K.; Jansen, L. E.; Van Duyne, G. D.; Vinogradov, S. A.; Lampson, M. A.; Black, B. E. *Science* **2015**, *348*, 699−703.
(37) Black, B. E.; Brock, M. A.; Bédard, S.; Woods, V. L.; Cleveland, D. W. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 5008−5013.
(38) Hamiche, A.; Shuaib, M. *Biochim. Biophys. Acta, Gene Regul. Mech.* **2012**, *1819*, 230−237.
(39) Mattiroli, F.; D'Arcy, S.; Luger, K. *EMBO Rep.* **2015**, *16*, 1454.
(40) Foltz, D. R.; Jansen, L. E.; Bailey, A. O.; Yates, J. R.; Bassett, E. A.; Wood, S.; Black, B. E.; Cleveland, D. W. *Cell* **2009**, *137*, 472−484.
(41) Dunleavy, E. M.; Roche, D.; Tagami, H.; Lacoste, N.; Ray-Gallet, D.; Nakamura, Y.; Daigo, Y.; Nakatani, Y.; Almouzni-Pettinotti, G. *Cell* **2009**, *137*, 485−497.
(42) Zasadzińska, E.; Barnhart-Dailey, M. C.; Kuich, P. H. J.; Foltz, D. R. *EMBO J.* **2013**, *32*, 2113−2124.
(43) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.
(44) Kato, H.; Zhou, B.-R.; Feng, H.; Bai, Y. *Cell Cycle* **2013**, *12*, 3133−3134.
(45) Kato, H.; Jiang, J.; Zhou, B.-R.; Rozendaal, M.; Feng, H.; Ghirlando, R.; Xiao, T. S.; Straight, A. F.; Bai, Y. *Science* **2013**, *340*, 1110−1113.
(46) Yu, Z.; Zhou, X.; Wang, W.; Deng, W.; Fang, J.; Hu, H.; Wang, Z.; Li, S.; Cui, L.; Shen, J.; Zhai, L.; Peng, S.; Wong, J.; Dong, S.; Yuan, Z.; Ou, G.; Zhang, X.; Xu, P.; Lou, J.; Yang, N.; Chen, P.; Xu, R.; Li, G. *Dev. Cell* **2015**, *32*, 68−81.
(47) Carroll, C. W.; Milks, K. J.; Straight, A. F. *J. Cell Biol.* **2010**, *189*, 1143−1155.
(48) Arimura, Y.; Shirayama, K.; Horikoshi, N.; Fujita, R.; Taguchi, H.; Kagawa, W.; Fukagawa, T.; Almouzni, G.; Kurumizaka, H. *Sci. Rep.* **2014**, *4*, 7115.
(49) Bassett, E. A.; DeNizio, J.; Barnhart-Dailey, M. C.; Panchenko, T.; Sekulic, N.; Rogers, D. J.; Foltz, D. R.; Black, B. E. *Dev. Cell* **2012**, *22*, 749−762.
(50) Logsdon, G. A.; Barrey, E. J.; Bassett, E. A.; DeNizio, J. E.; Guo, L. Y.; Panchenko, T.; Dawicki-McKenna, J. M.; Heun, P.; Black, B. E. *J. Cell Biol.* **2015**, *208*, 521−531.
(51) Quénet, D.; Dalal, Y. *Chromosome Res.* **2012**, *20*, 465−479.
(52) Tachiwana, H.; Müller, S.; Blümer, J.; Klare, K.; Musacchio, A.; Almouzni, G. *Cell Rep.* **2015**, *11*, 22−32.
(53) English, C. M.; Adkins, M. W.; Carson, J. J.; Churchill, M. E.; Tyler, J. K. *Cell* **2006**, *127*, 495−508.
(54) Mizuguchi, G.; Shen, X.; Landry, J.; Wu, W.-H.; Sen, S.; Wu, C. *Science* **2004**, *303*, 343−348.
(55) Mattiroli, F.; D'Arcy, S.; Luger, K. *EMBO Rep.* **2015**, *16*, 1454.
(56) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845.
(57) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(58) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.
(59) Joung, I. S.; Cheatham, T. E., III *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(60) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(61) Hess, B. *J. Chem. Theory Comput.* **2008**, *4*, 116−122.
(62) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
(63) Parrinello, M. *J. Appl. Phys.* **1981**, *52*, 7182.
(64) Amadei, A.; Ceruso, M. A.; Di Nola, A. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 419−424.
(65) Hess, B. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2002**, *65*, 031910.
(66) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 3352−3357.
(67) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.
(68) Parisi, G. *Phys. Rev. Lett.* **1983**, *50*, 1946.

DOI: 10.1021/jacs.6b05355
J. Am. Chem. Soc. 2016, 138, 13207−13218



PrePPI: A Structure Informed Proteome-wide Database of Protein–Protein Interactions

Donald Petrey [1†], Haiqing Zhao [1†], Stephen J Trudeau [1†], Diana Murray [1] and Barry Honig [1,2,3,4*]

*1 - **Department of Systems Biology,** Columbia University Irving Medical Center, New York, NY 10032, USA*
*2 - **Department of Biochemistry and Molecular Biophysics,** Columbia University Irving Medical Center, New York, NY 10032, USA*
*3 - **Department of Medicine,** Columbia University, New York, NY 10032, USA*
*4 - **Zuckerman Mind Brain and Behavior Institute,** Columbia University, New York, NY 10027, USA*

Correspondence to Barry Honig:1130 St. Nicholas Ave., Room 815, New York, NY 10032, USA.bh6@columbia.edu (B. Honig)*
https://doi.org/10.1016/j.jmb.2023.168052
Edited by Michael Sternberg

Abstract

We present an updated version of the Predicting Protein-Protein Interactions (PrePPI) webserver which predicts PPIs on a proteome-wide scale. PrePPI combines structural and non-structural evidence within a Bayesian framework to compute a likelihood ratio (LR) for essentially every possible pair of proteins in a proteome; the current database is for the human interactome. The structural modeling (SM) component is derived from template-based modeling and its application on a proteome-wide scale is enabled by a unique scoring function used to evaluate a putative complex. The updated version of PrePPI leverages AlphaFold structures that are parsed into individual domains. As has been demonstrated in earlier applications, PrePPI performs extremely well as measured by receiver operating characteristic curves derived from testing on *E. coli* and human protein–protein interaction (PPI) databases. A PrePPI database of ∼1.3 million human PPIs can be queried with a webserver application that comprises multiple functionalities for examining query proteins, template complexes, 3D models for predicted complexes, and related features (https://honiglab.c2b2.columbia.edu/PrePPI). PrePPI is a state-of-the-art resource that offers an unprecedented structure-informed view of the human interactome.

Introduction

The identification of proteins that interact with one another is a challenging problem of central importance in fundamental biology and in medicine. Protein-protein interactions (PPIs) is a widely used term which has multiple meanings. Two proteins can interact with one another directly either by forming a binary physical complex or by being in physical contact in the context of a multi-protein complex. Indirect interactions can include two proteins that are part of a complex, but are not in physical contact, or that are part of a pathway or network that mediates their interaction. Multiple experimental and computational tools are available to detect or predict PPIs, and their results are compiled in multiple databases. Here we report a new version of our Predicting Protein-Protein Interactions (PrePPI) database,[1–2] describe its unique features, and compare its performance to that of other databases. We also place PrePPI's prediction algorithm in the context of recent structure-based, co-evolution, and deep learning-based developments in the prediction of PPIs.

The key element of the PrePPI algorithm, which is summarized in Figure 1, is proteome-wide template-based modeling of PPIs, both direct and indirect. Not accounting for splice variants and posttranslational modifications, there are ~200 million possible non-redundant pairwise combinations of human proteins. However, since we consider full proteins as well as their individual domains, we need to examine ~4.55 billion pairwise interactions and, since we make multiple

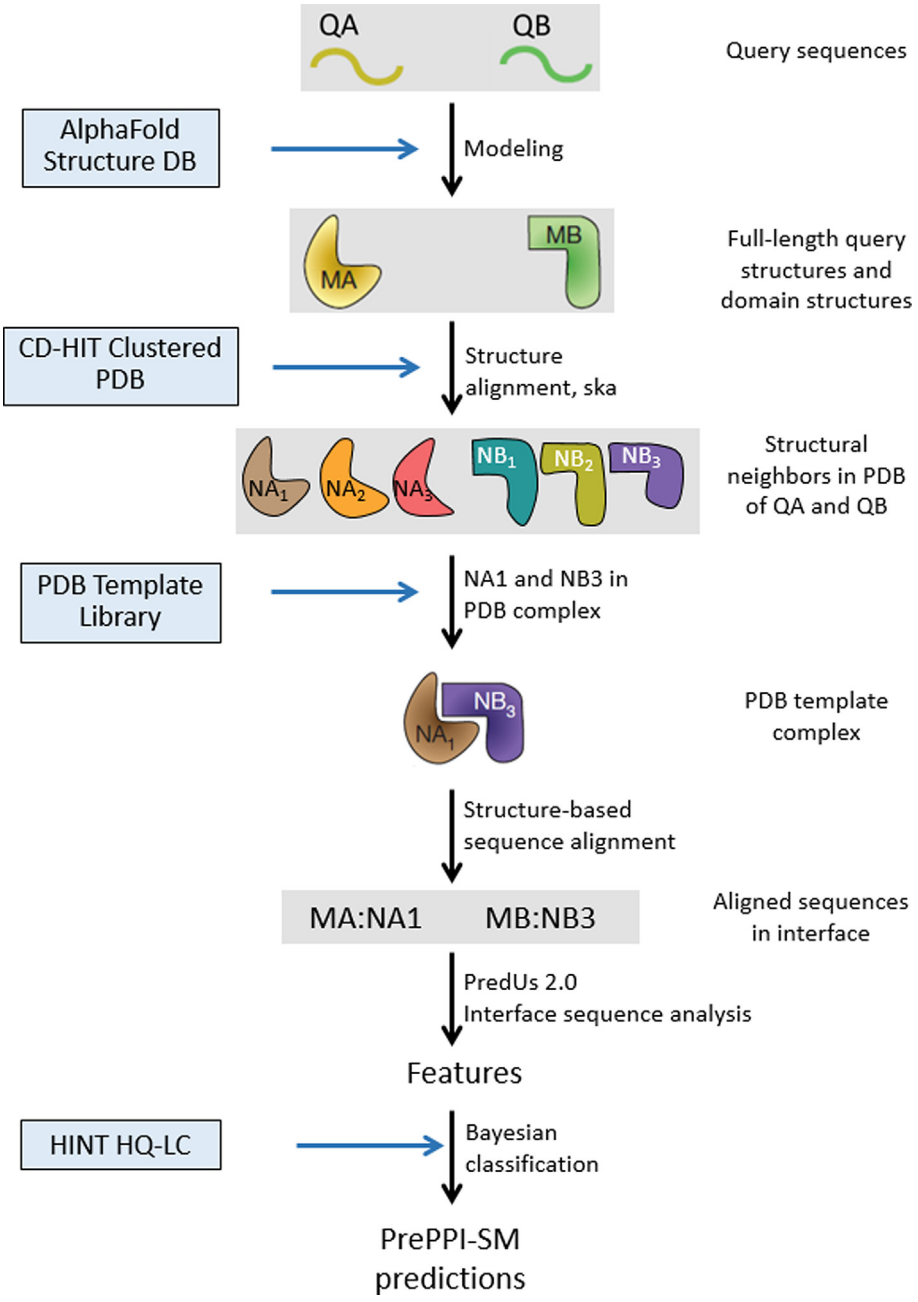


Figure 1. PrePPI's structural modeling (SM) pipeline: Structures for query proteins, QA and QB, are taken from the AlphaFold Protein Structure Database[13] and parsed into domains with definitions from the Conserved Domain Database (CDD) as MA and MB.[22] Structural neighbors in the PDB[3] for full length protein and domain structures with definitions from the Evolutionary Classification of Protein Domains (ECOD) database are obtained from the ska structural alignment program.[31] If structural neighbors of two query proteins appear together in a PDB complex, this structure defines a template, NA1:NB3, used to create a structure-based sequence alignment with which an interface for the query proteins, MA:MB, is evaluated based on the overlap of the query and template residues.[1] The interaction is then scored based on a number of features[1–2] and trained on the HINT HQ-LC database,[10] as the positive set, and a negative set described in Methods to produce a fully connected Bayesian network used to evaluate the model.

interaction models for each pair, the number of pairwise combinations evaluated is in the tens of billions (see Methods). PrePPI's ability to consider such a large number of potential PPIs is enabled by an efficient scoring function which is based on the similarity of the modeled interface to the interface of a known complex in the Protein Data Bank (PDB).[3] We highlight these points because it is important to distinguish our goals from standard template-based modeling. Furthermore, we are not necessarily trying to produce an accurate model of the complex as might be judged, for example, in the CAPRI (Critical Assessment of PRediction of Interactions) experiment[4] – although obviously a better model will produce a more reliable prediction. Rather, our hypothesis is that, in the derivation of a structural modeling score, our models are good enough to provide evidence that two proteins form a physical complex. Thus, a model that would score poorly according to CAPRI metrics might be reliable enough to provide a yes or no prediction as to whether two proteins interact and, in addition, produce a low-resolution structural pose for the interaction. As discussed below, PrePPI uses non-structural information as well. For example, if two proteins are co-expressed and have a good structural modeling (SM) score, the likelihood of an interaction, as given in PrePPI by a naïve Bayesian network, will increase. A PPI with low SM score but high non-structural score suggests that the interaction is indirect.

Testing and validating computational predictions is a complicated challenge since experimental databases themselves contain sources of uncertainty and the degree of overlap between them is still quite low in spite of the proliferation of observations from high-throughput screens. Moreover, they are often based on different definitions of PPIs. Mass spectrometry-derived databases (e.g. Bioplex 3.0[5]) focus on multi-protein complexes[6] while Y2H-based databases (e.g. HuRI[7]) focus on binary interactions. Among derived databases, the widely used STRING database[8] has a category for physical interactions but does not distinguish binary interactions from those in multi-protein complexes whereas databases such as APID[9] and HINT[10] include both direct and indirect interactions and attempt to distinquish between the two. As depicted in Table 1, overlap between these various databases is limited (see Methods for a description of each database). Of note, Interactome3D which contains PDB structures and high quality homology models is well-represented in most of the databases, but the HINT high-quality literature-curated database (HINT HQ-LC) contains the highest percentage of Interactome3D structures.

In earlier versions of PrePPI,[1–2] training was done on yeast PPIs and testing was done on human interactions, with the true positive dataset comprising PPIs with at least two literature references. No attempt was made at the time to train on datasets of binary physical interactions since PrePPI predicts both direct and indirect interactions. Here we have taken a more refined approach, training the structural modeling component of PrePPI on HINT HQ-LC human PPIs.[10]

In order to evaluate PrePPI's structure-based algorithm, we have used *Escherichia coli* K-12 (here *E. coli)* as a test organism and compared predictions from PrePPI's structural modeling component to predictions from the threading component of Threpp.[11] Technology closely related to Threpp powers the PEPPI server[12] which, like PrePPI, uses Bayesian statistics to integrate structural and non-structural information. But in contrast to the PrePPI, the PEPPI webserver allows a user to input only two protein sequences at a time while,

Table 1 Overlap among PPI databases: The number of overlapping entries among the databases denoted (see Methods) is listed for **A**. *E. coli* and **B**. Human.

A	Interactome3D	HINT HQ-LC	APID Level 2	STRING-Physical
Interactome3D	1,391			
HINT HQ-LC	1,092	1,675		
APID Level 2	381	363	3071	
STRING-Physical	396	651	2,322	10,577

B	Interactome3D	HINT HQ-LC	HINT HQ-Binary	APID Level 2	STRING-Physical	PrePPI-2016	HURI	BIOGRID-MV
Interactome3D	15,629							
HINT HQ-LC	8,639	15,598						
HINT HQ-Binary	11,761	15,598	119,526					
APID Level 2	9,092	8,098	102,130	154,955				
STRING-Physical	9,519	9,888	29,761	40,161	272,361			
PrePPI-2016	4,830	6,623	8,038	8,017	16,569	26,982		
HURI	1,875	1,107	34,743	33,578	6,335	695	39,060	
BIOGRID-MV	6,692	8,230	14,040	17,120	54,531	15,369	2,173	78,189

as described below, the PrePPI database of human PPIs contains about 200 million entries with the highest confidence predictions (~1.3 M) appearing in the online application that can be queried in multiple ways including, for example, inputting a single protein and outputting all predicted binding partners.

Compared to previous versions of PrePPI, in addition to improved training, features of the current version include the replacement of homology models with models from the AlphaFold Protein Structure Database[13] leading to increased structural coverage of the proteome, separate training of the structural modeling and non-structural components, a refined definition of PDB template complexes,[3] the implementation of a more accurate algorithm PredUs 2.0 for predicting interfacial residues,[14] and a website with expanded functionality. PrePPI is a unique resource that generates novel hypotheses for the existence of PPIs, both direct and indirect. Moreover, given the ongoing developments in the use of deep learning-based approaches to predict the structure of binary complexes, PrePPI predictions can be used as a starting point for the construction of accurate structural models.

Results

Testing on experimental databases

E. coli: We have chosen to test the SM score on *E. coli*, in part for comparison with Threpp[11] and in part to assess the applicability of our human-trained Bayesian network (see below) to another organism. PrePPI for *E. coli* was trained on human HINT HQ-LC[10] (see Methods). Table 2A presents area under the ROC curve (AUROC) values for the structural modeling component of PrePPI (PrePPI-SM) and the threading component of Threpp (Threpp-Threading)[11–12] for *E. coli* evaluated on three datasets: HINT HQ-LC and Interactome3D PPIs for *E. coli*, and GS-Threpp,[15] the gold standard data set of 763 PPIs on which Threpp was previously tested.[11] Both methods yield good results when tested on HINT HQ-LC (AUROC values 0.88 and 0.81 for PrePPI-SM and Threpp-Threading, respectively) and Interactome3D (AUROC values 0.95 and 0.85) but performance degrades (AUROC values 0.67 and 0.65) on GS-Threpp. PrePPI-SM performs quite well on HINT HQ-LC and performance improves on Interactome3D which is comprised of PDB complexes or close homologs.[16] As can be seen in Table 1A, HINT HQ-LC has a large intersection with Interactome3D (65%). The slight difference in performance may arise if some of the interactions in HINT HQ-LC are not readily homology-modeled. Overall, the PrePPI-SM results are somewhat better than those obtained with Threpp-Threading but it is reassuring that two different structure-based methods yield very similar performance and, in particular, that a proteome-wide method such as PrePPI is of comparable accuracy to a method that uses a more complex and computationally intensive scoring function to evaluate structural models.

Human: Table 2B presents AUROC values for PrePPI-SM and PrePPI-Total, where the latter corresponds to the predicted score with all sources of evidence (Figure 1), with testing on HINT HQ-LC and the high confidence set we assembled in 2016, PrePPI-2016.[2] PrePPI-SM performs very well on HINT HQ-LC (AUC = 0.83) but performance degrades on PrePPI-2016 (AUC = 0.73). We attribute the difference to the fact that HINT HQ-LC was designed to encompass experimentally observed direct PPIs and, thus, has significant overlap (56%) with Interactome3D[16] (Table 1B) while PrePPI-2016 contains many indirect interactions (19% overlap with Interactome3D). Consistent with this explanation, the difference in performance between the use of just structural evidence or the combination of structural and non-structural evidence for testing on HINT HQ-LC (AUROC = 0.83 for PrePPI-SM and 0.77 for PrePPI-Total) is small, whereas the AUROC for testing on the PrePPI-2016 set increases from 0.73 for PrePPI-SM to 0.89 for PrePPI-Total, indicating that PrePPI-Total successfully captures both structural and non-structural evidence.

Table S1 contains AUROC values for PrePPI-Total tested on a number of PPI databases. The values vary over a wide range which appears to reflect underlying differences in the databases as delineated in Table 1. As summarized in Methods, HURI,[7] HINT HQ-Binary[10] and APID Level 2[9] contain many Y2H results, STRING-Physical[17] contains many direct and indirect physical

Table 2 Area under ROC curve, AUROC, for different test sets. **A.** *E coli.* The performance of PrePPI-SM compared to that of Threpp-Threading, both tested on Interactome3D, Hint HQ-LC and GS-Threpp. **B.** *Human.* The performance of PrePPI-SM and PrePPI-total tested on Hint HQ-LC and the PrePPI 2016 high confidence set (PrePPI-2016).

A	HINT HQ-LC	Interactome3D	GS-Threpp
PrePPI-SM	0.88	0.95	0.67
Threpp-Threading	0.81	0.85	0.65

B	HINT HQ-LC	PrePPI-2016
PrePPI-SM	0.83	0.73
PrePPI-Total	0.77	0.89

interactions, and BioGRID-MV[18] infers PPIs from a large range of experimental methods. HINT HQ-LC is derived from binary interactions that have at least two literature references and, in that sense, is most closely related to PrePPI-2016. Agreement between PrePPI and HURI is quite limited (see Luck et al.[7] for a discussion of HURI's overlap with other databases). Of course, it is impossible to know how many predicted PPIs that do not appear in any database are actually true positives. Indeed PrePPI's goal is to discover PPIs that do not appear in known databases. Based on experimental tests and applications summarized in the Discussion, PrePPI has already proved to be a reliable source of novel PPIs.

To place PrePPI predictions in the context of deep learning approaches, we compared PrePPI performance to that of D-SCRIPT,[19] a proteome-wide method for predicting physical interactions between two proteins given just their sequences. Similar to PrePPI, D-SCRIPT was trained on human PPIs and predicts PPIs for both human and *E. coli*, however training and testing were performed with PPIs from the STRING database[17] whereas PrePPI used HINT HQ-LC[10] (see comparisons in Table 1A and B). In spite of the differences in training and testing sets, the performance, as judged by AUROC values, is similar for both *E. coli* (PrePPI-SM: 0.88, D-SCRIPT: 0.86) and human (PrePPI-SM: 0.83, D-SCRIPT: 0.83) PPIs. Given the low overlap between the HINT HQ-LC and STRING-Physical databases, the strong performance of both methods suggests they are highly complementary, not only in methodological terms but also in the type of information they encompass.

The PrePPI database: The full PrePPI database contains predictions for ~200 million PPIs. Even though interaction models are evaluated for a protein and its constituent domains, only the highest scoring interaction for a given protein pair is included in the database. Hence, the set of 200 million non-redundant PPIs corresponds to near total coverage of all possible interactions among ~20 K proteins. The online database contains about 1.3 M human PPIs of which about 370 K represent predictions of direct physical interactons. PPIs that appear in the online database either are associated with an FPR < 0.005 (LR > 379) or have the maximum value of LR(SM) or LR(protein-peptide) > 100. Our experience has been that interactions that meet this latter criterion constitute high-confidence physical interactions and, indeed, are associated with an FPR < 0.001 when tested on the structure-rich HINT HQ-LC database.

PrePPI website (https://honiglab.c2b2.columbia.edu/PrePPI/): When a user inputs a UniProt ID or gene name for a query protein, the website returns several features of the protein and its predicted interactors: 1) the names and functional information for the query protein derived from UniProt; 2) the sequence of the full-length query protein as well as its domains, all of which can be viewed in a protein-centric structure viewer; 3) a list of PrePPI-predicted interactors of the query protein and associated scores for the features incorporated in the PrePPI algorithm, and, if they exist for a given PPI, links to external databases that compile interactions based on experiments and literature; 4) an interaction-centric structure viewer that shows the 3D model for a given PPI and, depending on selections by the user, the template PDB complex and the structure superposition of the query structures on the template (Figure 1); 5) functional annotations for the query protein, derived from gene set enrichment analysis of the protein's interactors ranked according to the PrePPI-Total score[2]; 6) annotations of the full-length query protein sequence for disordered regions[20]; and 7) annotations of the full-length query protein sequence for interfacial residues as predicted by PredUs 2.0[14] that is used in the PrePPI-SM scoring function (Figure 1).

Discussion

The PrePPI database was first reported in 2012[1] and updated in 2016.[2] Its unique features include a fast structure-based scoring function that enables proteome-wide protein–protein interface evaluation and the integration of structural and non-structural evidence for an interaction. The current version of PrePPI has been improved in a number of ways: 1) Most notably, our in-house homology model database has been replaced with structures from the AlphaFold Protein Structure Database[13] for individual proteins and their domains as annotated by the Conserved Domain Database (CDD).[21] As explained in Methods, use of the AF/CDD database requires the scoring tens of billions of interaction models. This scoring takes about a day using ~2000 CPU processors. 2) The training of structure-based versus non-structural evidence is performed separately. Specifically, the structure-informed predictions are trained with the HINT HQ-LC database[10] while non-structural features are derived as implemented previously[2] and trained on databases with a predominance of non-structural information. 3) The method to extract non-crystallographic protein–protein interfaces from the PDB has been revised. 4) A more accurate algorithm, PredUs 2.0, was implemented for predicting interfacial residues on protein surfaces.[14] 5) New website features are as described above.

We are not aware of any structure-informed database comparable in scope to PrePPI. Many of its predictions have not been previously observed since use of 3D structure information, especially in matching protein structures to PPI template complexes from the PDB, identifies many interactions that would be undetectable with

sequence-based methods. PrePPI performance is comparable to that of high-throughput experimental methods.[1–2] Moreover, experimental validation has already confirmed the reliability of many novel predictions: 1) In the original PrePPI paper,[1] 17 out of 21 predictions were confirmed with co-IP assays; 2) In our study of virus/human interactions with the P-HIPSTer database, which is based on the PrePPI pipeline,[22] PrePPI predictions yielded a 76% precision as judged by co-IP experiments; 3) PrePPI is a central feature in the OncoSig algorithm that generated a lung cancer adenocarcinoma (LUAD) signaling PPI network for KRAS that recapitulated published KRAS biology and identified novel proteins synthetic lethal with an oncogenic mutated form of KRAS that is constitutively activated;18 of 21 were validated in 3D spheroid models for LUAD.[23] Thus, based on results in a wide range of contexts, PrePPI predictions are associated with a precision of ~75–80%.

Of course, not all PrePPI predictions are correct but, as highlighted in the previous paragraph, they appear sufficiently accurate to generate hypotheses that drive biological discovery. Moreover, for direct binary PPIs, a model that appears in the database can be used as a basis for lower throughput approaches such as protein–protein docking or deep learning algorithms such as AlphaFold multimer[24] which likely generate models that are more accurate than those in PrePPI. PrePPI predictions for non-direct interactions also provide valuable information by identifying pairs of proteins that might be present in multi-protein complexes and, moreover, PrePPI predictions can be used to identify all proteins that are in physical contact in such a complex.[2] PrePPI predictions can also be used in the construction of PPI networks that comprise both direct and indirect interactions and, when combined with features based on context-specific gene expression or knockout screens, can provide insight into dysregulation of cellular signaling as demonstrated with the KRAS-centered OncoSig network for LUAD.[23]

Given the continuous developments in structure determination and sequence analysis, PrePPI will continue to evolve and to incorporate new technologies. One possibility is to leverage the proteome-wide, complementary approaches of PrePPI and D-SCRIPT[19] and integrate the interface predictions from both as features in an enhanced PPI prediction algorithm. More computationally intensive methods such as ECLAIR[25] can be used to filter PrePPI predictions thus improving their accuracy. While such methodological advances are in development, the current version of PrePPI will be applied to multiple proteomes and to cross-species interactions as implemented in our P-HIPSTer database.[22] In summary, we believe that PrePPI constitutes a unique resource that will continue to find applications in multiple areas of biomedical science.

Methods

Training the SM score

Extracting biological interfaces from the PDB: All possible PDB complexes, regardless of source organism, are considered. The quaternary structure of a PDB file frequently does not represent the biologically relevant quaternary structure[26] but will be represented by one of the "biological assemblies" contained in the PDB file. The biological assemblies are specified in the "REMARK 350" lines of the PDB file and contain a set of geometric transformations ("BIOMT" records). A given biological assembly is constructed by applying the transformations defined for that assembly to the set of chains in the PDB file. To define template interface contacts, we construct three-dimensional models of each biological assembly using the associated transformations. A contact between any pair of chains in a biological assembly is defined when two heavy atoms across the interface are within 6 Å of each other. The union of these contacts from all biological assemblies for each pair of chains comprises the interface for those chains and is used to evaluate structure-based predictions as described in the following sections. ~200 K PDB structures, each of which contain, on average, several bioassemblies, are used to construct interfaces.

Model Building: Sequences for the human and *E. coli* K12 proteomes are taken from the UniProt defined reference proteomes with one representative protein per gene (Proteome IDs UP000005640 and UP000000625, respectively).[27] As we recently described,[28] each full-length sequence is broken up into individual domains corresponding to those defined in the CDD.[21] Three-dimensional models for each full-length protein are taken from the AlphaFold Protein Structure Database[13] with models for individual domains extracted from the model of the full-length protein. This generates models databases that structurally represent 1) 20,251 human proteins with 20,251 full-length sequence models and 69,678 CDD domain models, and 2) 4,463 *E. coli* proteins with 4,463 full-length sequence models and 7,713 CDD domain models.

Interaction Model Construction: Sequences for every protein chain in the PDB are downloaded from the PDB web site.[3] The sequences are clustered at a sequence identity cutoff of 60% using the program CD-HIT[29] to form PDB sequence clusters, and a representative for each cluster is defined as the longest sequence in the cluster. The structures corresponding to a PDB sequence cluster include the full-length PDB structures and their constituent domains as defined by the Evolutionary Classification of Domains (ECOD) database.[31] For a given query protein, the sequences for its associated models are matched to PDB sequence clusters and the query models are structurally aligned

to the PDB structure for the representative of the corresponding cluster. The quality of the structure alignment is scored using the Protein Structural Distance (PSD) calculated from the program ska.[30] Of note, in practice, ska alignments involve protein structures with at least three secondary structure elements so that, beyond PrePPI's use of sequence orthology as an evidence source, PrePPI typically does not predict interactions involving a single α-helix to a structured domain. If a query model aligns with a PSD < 0.6 to the structure of the representative sequence of a PDB cluster or its domains as defined by ECOD,[31] the query model is further aligned to all of the cluster structures. PDB structures with PSD < 0.6 are kept as structural neighbors of the query model. Whenever the structures for the structural neighbors of two query proteins appear together in a PDB complex (as defined above), we call this complex a "template" for an interaction of the query proteins. In practice, we never create a three-dimensional interaction model, rather the structure-based sequence alignments between the query protein models and the identified interaction model template chains are used to derive properties of the interaction: the quality of the alignment itself; the extent that residues of the query proteins align to interfacial residues in the template; and the extent to which residues predicted to be interfacial in the query proteins align to interfacial residues in the template.[1] Predicted interfacial residues are obtained from our program PredUs 2.0.[14] This scoring avoids the need to explicitly calculate pairwise properties while preserving context-specific information for the template complex and enables rapid evaluation of interaction models from among billions of possible pairwise query combinations.

Given that the full length protein and multiple domains are used for each protein and multiple models are tested for each of the 90 K human query sequences, tens of billions of interaction models must be evaluated. Each model is evaluated using a scoring function derived from a Bayesian network based on features as summarized above and reported previously.[2] Training of the Bayesian network is based on training sets as described below. For a given protein pair, the highest scoring interaction, whether it is between two full length proteins or between two domains, is chosen for that PPI, leading to a non-redundant set of about 200 million scored predictions.

True positive data sets: The most obvious training set for direct interactions is the PDB[3] but it contains a relatively limited number of entries for complexes in a given proteome and redundancies further limit this number. Instead, we have preferred to use the HINT high-quality literature-curated database, HINT HQ-LC,[10] which appears to be the best source for direct physical interactions and currently has 16 K entries for human and 1,753 for *E. coli*.

We have used a number of databases to calculate ROC curves. The size of these databses and the overlap between them appear in Table 1. They include:

Interactome3D[16]: PDB structures and easily constructed homology models.

HINT high-quality literature-curated (HINT HQ-LC)[10]: Experimentally observed binary PPIs with at least two literature references.

APID Level 2[9]: Interactions experimentally observed by at least 1 binary method.

STRING-Physical[8]: Direct and indirect PPIs in the same complex with experimental evidence.

BioGRID-MV[18]: PPIs curated from both high-throughput datasets and individual focused studies that are validated by multiple experiments.

HURI[7]: Binary PPIs validated by three variations of the Y2H assay.

Overall, the lack of overlap among different databases highlights questions about how they are used/chosen in the training of computational methods, especially for those focused on direct interactions. Our decision to train the structural component on a different true positive set than that used for the non-structural component is an attempt to address this issue. For both human and *E. coli*, HINT HQ-LC has significant overlap with Interactome3D consistent with its focus on direct interactions.

True negative data set: The negative set used in training and testing consists of all possible human PPIs minus the union of PPIs that appear at any level of confidence in the databases listed in the previous section. The treatment of every interaction for which there is no evidence as a true negative obviously diminishes apparent performance. But our experience has been that, as opposed to precision/recall curves, ROC curves are not significantly affected by the size of the negative set. We have confirmed this behavior by changing the size of the negative set to be 10 times the size of the positive set and found that this has essentially no effect on the various ROC curve statistics. Specifically, the values in Table S1 are identical using either negative set. In addition, Figure S2 shows complete overlap between between ROC curves using both negative sets as tested on two different data sets.

Training the non-structural score

As reported previously, in addition to structural evidence, PrePPI uses a number of non-structural features including partner redundancy, GO (gene ontology) annotation, sequence orthology, and phylogenetic profile. Details about the calculation and training of non-structural contributions are described in our 2016 publication[2] and will not be repeated here. Briefly, the true positive set was taken from multiple databases with the requirement that a PPI be identified in two independent literature

references and no attempt was made to distinguish direct physical from non-direct interactions.

CRediT authorship contribution statement

Donald Petrey: Conceptualization, Software, Validation, Methodology. **Haiqing Zhao:** Formal analysis, Investigation, Data curation. **Stephen J Trudeau:** Formal analysis, Investigation, Software. **Diana Murray:** Conceptualization, Writing – original draft, Formal analysis, Investigation. **Barry Honig:** Conceptualization, Writing – original draft, Funding acquisition.

DATA AVAILABILITY

Data will be made available on request.

Acknowledgements

This work was supported by the National Institute of Health (grant R35-GM139585 (BH), grants T32-GM008224 and T32-GM145440 (SJT)).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.168052.

Received 6 January 2023;
Accepted 10 March 2023;
Available online 17 March 2023

Keywords:
protein-protein interactions;
database;
alphafold models;
structural modeling;
non-structural evidence

† Co-first author.

References

1. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556–560.
2. Garzon, J.I., Deng, L., Murray, D., Shapira, S., Petrey, D., Honig, B., (2016). A computational interactome and functional annotation for the human proteome. *Elife*, 5.
3. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
4. Lensink, M.F., Brysbaert, G., Mauri, T., Nadzirin, N., Velankar, S., Chaleil, R.A.G., (2021). Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins* **89**, 1800–1823.
5. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040.e28.
6. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M. P., Szpyt, J., (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440.
7. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., (2020). A reference map of the human binary protein interactome. *Nature* **580**, 402–408.
8. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612.
9. Alonso-Lopez, D., Campos-Laborie, F.J., Gutierrez, M.A., Lambourne, L., Calderwood, M.A., Vidal, M., (2019). APID database: redefining protein-protein interaction experimental evidences and binary interactomes. *Database (Oxford)* **2019**
10. Das, J., Yu, H., (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92.
11. Gong, W., Guerler, A., Zhang, C., Warner, E., Li, C., Zhang, Y., (2021). Integrating Multimeric Threading With High-throughput Experiments for Structural Interactome of Escherichia coli. *J. Mol. Biol.* **433**, 166944
12. Bell, E.W., Schwartz, J.H., Freddolino, P.L., Zhang, Y., (2022). PEPPI: Whole-proteome Protein-protein Interaction Prediction through Structure and Sequence Similarity, Functional Association, and Machine Learning. *J. Mol. Biol.* **434**, 167530
13. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
14. Hwang, H., Petrey, D., Honig, B., (2016). A hybrid method for protein-protein interface prediction. *Protein Sci.* **25**, 159–165.
15. Hu, P., Janga, S.C., Babu, M., Diaz-Mejia, J.J., Butland, G., Yang, W., (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96.
16. Mosca, R., Ceol, A., Aloy, P., (2013). Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53.
17. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815.
18. Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541.
19. Sledzieski, S., Singh, R., Cowen, L., Berger, B., (2021). D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst.* **12**, 969–982.e6.

20. Dosztanyi, Z., (2018). Prediction of protein disorder based on IUPred. *Protein Sci.* **27**, 331–340.
21. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229.
22. Lasso, G., Mayer, S.V., Winkelmann, E.R., Chu, T., Elliot, O., Patino-Galindo, J.A., (2019). A Structure-Informed Atlas of Human-Virus Interactions. *Cell* **178**, 1526–1541. e16.
23. Broyde, J., Simpson, D.R., Murray, D., Paull, E.O., Chu, B. W., Tagore, S., (2021). Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat. Biotechnol.* **39**, 215–224.
24. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., et al., (2022). Protein complex prediction with AlphaFold-Multimer. *bioRxiv*. 2021.10.04.463034.
25. Meyer, M.J., Beltran, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., (2018). Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107–114.
26. Krissinel, E., Henrick, K., (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797.
27. UniProt, C., (2022). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*.
28. Trudeau, S.J., Hwang, H., Mathur, D., Begum, K., Petrey, D., Murray, D., et al., (2023). A structure- and chemical similarity-informed database of predicted protein compound interactions. *Protein Sci.*, e4594. https://doi.org/10.1002/pro.4594.
29. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.
30. Yang, A.S., Honig, B., (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301**, 665–678.
31. Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., Kim, B.H., Grishin, N.V., (2014). ECOD, An evolutionary classification of protein domains. *PLoS Comput Biol* **10**, (12) e1003926