Using Mathematical Tools to Analyze Chromatin Structures

Ondrej Maxian

DIMACS REU 2017

Mentor: Dr. Wilma Olson

July 28, 2017

Abstract:
The packing of DNA into nucleosome core particles (NCPs) is vital to the field of epigenetics and gene expression. Until now, the packing of NCPs has been studied physically via X-rays and cryogenic electron microscopy (cryo-EM), with few to little mathematical tools available. We here develop tools to quantitatively evaluate the amount of interaction between two nucleosomes on their faces and sides, as well as to give biological meaning to the regions of interaction. We apply our methods to X-ray crystal and cryo-EM structures, in the process confirming much of the existing observations while developing new insights into some atypical structures. We also deploy our algorithm to analyze configuration outputs from Markov Chain Monte Carlo (MCMC) simulations, which we found to closely resemble those from cryo-EM and X-ray.

I. Introduction

The field of epigenetics is concerned with how gene expression changes without modification of genetic code. One subfield of epigenetics deals with the packing of DNA, which is thought to control, among other processes, the ability of enzymes and proteins to transcribe DNA into messenger RNA and therefore translate mRNA into protein. Since DNA is usually found wrapped around core histone proteins to form nucleosome core particles (NCPs), which are in turn linked together with DNA to form chromatin fibers, it is important for the field of epigenetics to understand how NCPs pack together in chromatin fibers.

Unlike in DNA analysis, where motion between adjacent base pairs is generally small as defined by six narrowly distributed rigid body parameters, nucleosomes show large variations in these same parameters, which makes analysis of their structures all the more challenging. Any method developed to study the interaction of nucleosomes must consider the position of any nucleosome in a configuration with respect to any other one, and that method must also be able to translate so that results of biological significance can be obtained.

We here quantify nucleosomal interaction by defining polygons to represent the face and side of a nucleosome, which we treat as a modified cylinder. Then, given two nucleosomes, we compute the overlap between their faces and sides by projecting the respective polygons onto a common plane. We map the direction of one nucleosome with respect to another to a biological region on the nucleosome, which allows us to make inferences about regions of contact in X-ray crystal structures, and chromatin fibers obtained from both cryogenic electron microscopy (cryo-EM) and Markov Chain Monte Carlo (MCMC) simulations.

II. Building the model of the nucleosome as a simplified polyhedron

We model the nucleosome as a polygonal cylinder with a height of 44 Å in the normal direction. The coordinates of each atom that makes up the nucleosome core particle can be obtained from the Protein Data Bank (PDB, structure label 1kx5, [1]). Using the centers of the base pairs of the DNA that surrounds the nucleosome, a reference frame and origin can be computed for the NCP as follows: first, the origin is found by finding the point with the total least squares minimum distance to all of the base pair centers. A line going from this origin to the nucleosome dyad defines the first axis. The second axis, the normal vector, is then found by minimizing the sum of the dot products of the normal with all of the vectors from the origin to a base pair center. The center is then readjusted so that the first two vectors are orthogonal, and finally the third axis is computed via the cross product of the first two.

The reference frame consists of three orthonormal vectors: $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{n}$, where the $\boldsymbol{u}$ axis always goes from the origin of the coordinate system to the dyad, the $\boldsymbol{v}$ axis is another vector on the central face of the nucleosome, and the $\boldsymbol{n}$ axis is the vector normal to the central face of the nucleosome.

Now that the reference frame has been computed, the coordinates of each phosphorous atom on the DNA with respect to that reference frame can be calculated. Denoting $\boldsymbol{p_i}$ as a point in PDB coordinates, its coordinates with respect to the new reference frame are given by

$$\boldsymbol{p_i'} = T^*(\boldsymbol{p_i} - \boldsymbol{o_p})$$

Where $T = (\boldsymbol{u} \quad \boldsymbol{v} \quad \boldsymbol{n})$, $\boldsymbol{o_p}$ is the origin in PDB coordinates, and $\boldsymbol{p_i'}$ is the coordinates of the point with respect to the new reference frame.

Now, we consider every tenth phosphorous atom on the leading stand of the DNA (i.e. the strand used to designate the DNA sequence), beginning 35 base pairs away from the dyad (if the dyad

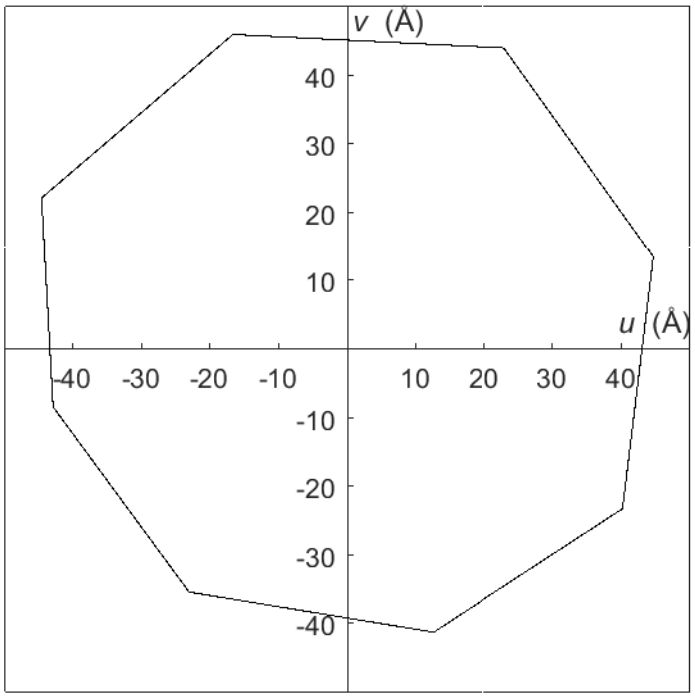

Figure 1: The polygon that defines the face of the nucleosome.

has number 0, the exact base pair numbers are then –35, –25, –15, –5, +5, +15, +25, +35). Projecting these eight atoms onto the plane of the nucleosome face by removing their normal coordinates allows us to generate an octagon that includes phosphorous atoms from around the entire nucleosome. This octagon is shown in Figure 1.

An important observation here is that this polygon is constant for every NCP. That is, if a vertex is given by $a\boldsymbol{u} + b\boldsymbol{v}$ in the reference frame generated from PDB coordinates, it will also be given by $a\boldsymbol{u} + b\boldsymbol{v}$ in any reference frame, since we can simply superimpose the 1kx5 structure onto any orthonormal reference frame that represents a nucleosome core particle without modification.

The procedure to compute the polygon can also be generalized to any structure in order to accurately compute overlaps in X-ray crystals. Given a reference frame and list of phosphate groups for the structure, we begin with the phosphorous atoms –35, –25, –15, –5, +5, +15, +25,

and +35 base pairs away from the dyad on the leading strand. By projecting their coordinates onto the $\boldsymbol{n} = 0$ plane of the structure, we generate a polygon that represents the structure face. We use this polygon, rather than the one for NCPs, to find overlap in these modified structures.

III. Finding face to face overlap

A. Finding a common plane

We consider the problem of finding the overlap between the faces of two nucleosomes. Each nucleosome can now be represented by the polygon in its $\boldsymbol{n} = 0$ plane, and the problem has been reduced to computing the overlap between polygons in three-dimensional space. In order to compute overlap in the two-dimensional sense, the polygons must be projected onto a common plane, and we use the "mid-frame" approach from DNA base pairs here [2]. The mid-frame approach takes in two sets of reference frames and origins and generates vectors $\boldsymbol{u_m}, \boldsymbol{v_m}, \boldsymbol{n_m}$and origin $\boldsymbol{o_m}$ that form a basis and origin for the mid-frame space (see appendix A for more information about computing the mid-frame). Given a set of vectors describing the polygon in one nucleosome basis $\boldsymbol{P}$, we can find their coordinates with respect to the mid-frame by applying two coordinate transformations:

$$\boldsymbol{P_m} = T_m^*\big((T\boldsymbol{P} + \boldsymbol{o}) - \boldsymbol{o_m}\big) \tag{1}$$

Where T is the matrix $T = (\boldsymbol{u} \quad \boldsymbol{v} \quad \boldsymbol{n})$ corresponding to the nucleosome and $\boldsymbol{o}$ is its origin. In this formulation, the matrix T transforms $\boldsymbol{P}$ from the nucleosome basis to the standard $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ basis, with translation by the origin $\boldsymbol{o}$. Next these points are translated so that the new origin is $\boldsymbol{o_m}$ and are then transformed by the conjugate transpose of T_m, T_m^*, from the standard basis to the mid-frame basis. The polygon at this point is not necessarily flat and may have nonzero normal coordinates in the mid-frame. To compute two dimensional overlap, the polygon $\boldsymbol{P_m}$ is projected onto the $\boldsymbol{n_m} = 0$ plane by setting the normal coordinates to zero and retaining the $\boldsymbol{u_m}$ and $\boldsymbol{v_m}$

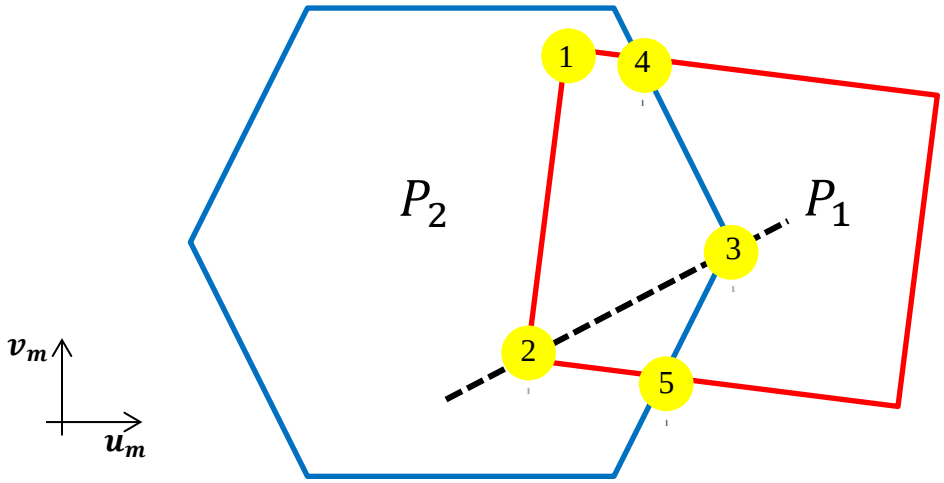

Figure 2: Computing the intersection of two polygons. Points 1 and 2 are vertices of P_1 that are inside P_2, point 3 is a vertex on P_2 that is inside P_1, and points 4 and 5 are where two edges of the shapes intersect. Arranging the points in a clockwise fashion allows the area of overlap to be computed from the area of the polygon of intersection.

coordinates. We do this for both polygons, so that now we are left trying to compute the overlap of two polygons in the $\boldsymbol{n_m} = 0$ plane.

B. Computing polygon overlap

It is first important to note that a linear map preserves convex combinations. Since the polygon in the nucleosome basis (Figure 1) is convex, its coordinates in the mid-frame also form a convex set. Since this set is convex in all three coordinates, $\boldsymbol{u_m}, \boldsymbol{v_m}, \boldsymbol{n_m}$, the projection of the polygon onto the $\boldsymbol{n_m} = 0$ plane is likewise a convex set. Therefore, the two polygons whose overlap is to be computed are convex. Their non-empty intersection, a subset of each set of polygon projections, is therefore convex. So computing the area of overlap is as simple as computing the area of a convex polygon that defines the overlap region.

Denote the polygons as P_1 and P_2. The set of vertices of the intersection of P_1 and P_2 is made up of vertices on P_1 that are inside P_2, vertices on P_2 that are inside P_1, and points where two edges of P_1 and P_2 intersect (see Figure 2). Denote this set of vertices as P_C. We order the set P_C in a clockwise fashion, cutting the shape along the line that goes through the vertices with minimum and maximum $\boldsymbol{u_m}$ coordinates (i.e. the dotted line in Figure 2). The points above the line are arranged by increasing $\boldsymbol{u_m}$ coordinate and ones below it by decreasing $\boldsymbol{u_m}$ coordinate. We then

append the list together (in Figure 2 the list is ordered 2-1-4-3-5) to obtain P_C in clockwise order. We can then compute the area of the overlap region by breaking it into triangles. Finally, we define an overlap coefficient, $\phi = \frac{A_{overlap}}{A_{polygon}}$, as a measure of the amount of overlap between the two nucleosomes ($\phi = 1$ means the nucleosomes share a common face, $\phi = 0$ means the nucleosomes have no overlap on their faces).

IV. Computing side to side overlap

When computing face to face overlap, the polygon is well defined in two dimensional space because the normal vector to the plane of the nucleosome is known. However, when computing the amount of side to side overlap between two nucleosomes, all that is known is that the vector normal to the polygon $\boldsymbol{P}$ that defines the face of the nucleosome must be in the plane that we use to compute side overlap. The direction of the plane that defines the side of the nucleosome is unknown. To visualize this, consider two nucleosomes at the corners of a three dimensional box. The side overlap between them will be observed if one looks from one corner to the other, not down any set axis. We therefore need to define an axis that looks from one corner to the other, or that can adjust so that we will obtain the maximum possible side overlap between two nucleosomes.

Figure 3 depicts our solution to this problem. We begin by drawing a line between the two nucleosome origins, $\boldsymbol{\ell} = \boldsymbol{o_B} - \boldsymbol{o_A}$ and computing its coordinates in each basis by multiplying by the appropriate linear operator. For example, to compute the coordinates of $\boldsymbol{\ell}$ in the basis of nucleosome A, we have $\boldsymbol{\ell_A} = T_A^* \boldsymbol{\ell}$, where T_A^* is the matrix whose *rows* are the basis vectors of nucleosome A. Next we project $\boldsymbol{\ell}$ onto the polygon defining the nucleosome face, $\boldsymbol{P}$, by removing the component of $\boldsymbol{\ell_A}$ that is in the direction normal to the face of the nucleosome.

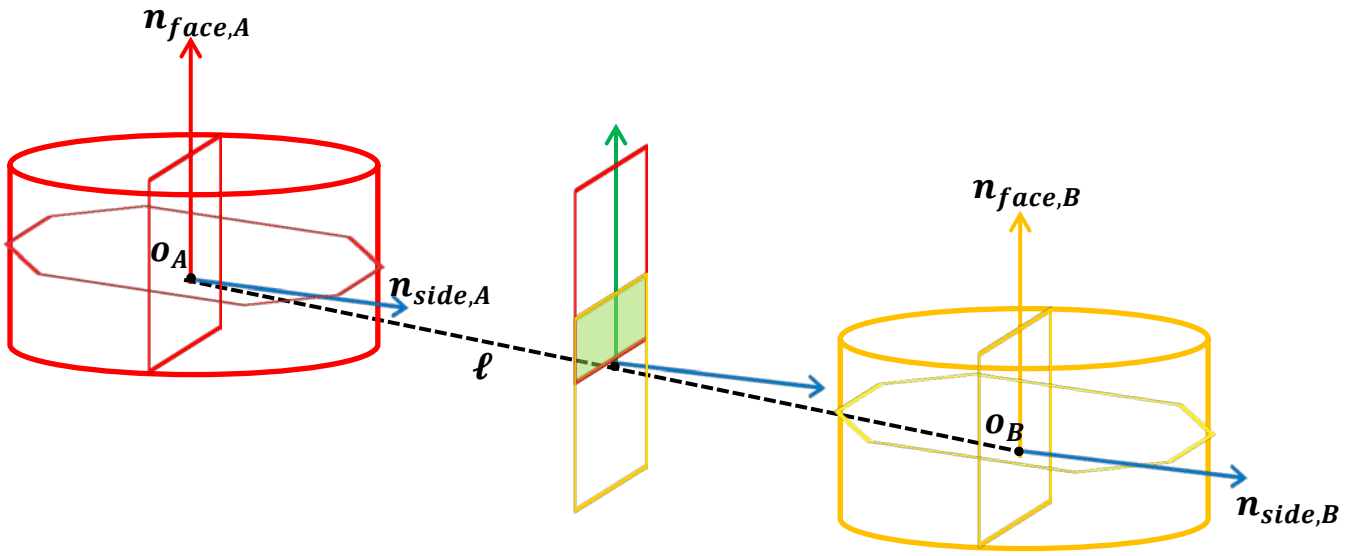

Figure 3: Computing side overlap. The side normal vectors shown in blue are computed for each nucleosome by eliminating the component of the line $\boldsymbol{\ell}$ that is normal to the nucleosome face. Then the mid-frame approach is used to obtain the "optimal" side overlap.

Denote this new vector, which is normal to the "side plane" of the nucleosome by $\boldsymbol{\ell}'_{\boldsymbol{A}}$. We can compute its absolute coordinates by transforming $\boldsymbol{\ell}'_{\boldsymbol{A}}$ back to the standard basis: $\boldsymbol{n_{side}} = T_A \boldsymbol{\ell}'_{\boldsymbol{A}}$. In this way, we arrange the nucleosomes so that the normal vector to the side plane points in the direction that generates maximal side overlap without overestimation (the "corners" of the box in the prior example).

As with the face overlap calculation, we impose a shape onto the side plane and use the mid-frame approach to calculate overlap between the two shapes. Unlike in face overlap, we use a rectangle, $\boldsymbol{R}$, for side overlap (see Figure 3). According to known measurements for the nucleosome core particle, this rectangle has a height of about 44 Å (in the direction normal to the nucleosome face) and a width of 90 Å (i.e. the radius of the nucleosome cylinder is 45 Å) [1]. We now have a convex polygon, $\boldsymbol{R}$, superimposed onto a coordinate frame that has a planar vector, $\boldsymbol{n_{face}}$, and a normal vector, $\boldsymbol{n_{side}}$. We can obtain a third vector (also in the plane of the rectangle) by taking the cross product $\boldsymbol{n_{face}} \times \boldsymbol{n_{side}}$. Now each shape has a three vector orthonormal basis associated with it. We proceed with the mid-frame approach (exactly the same as in the face overlap case) to compute the overlap between the two rectangles. This gives the side to side overlap from the optimal direction.

V. Using the direction of overlap to map to biological regions

Consider the vectors $\boldsymbol{n_{side,A}}$ and $\boldsymbol{n_{side,B}}$ as shown in Figure 3. $\boldsymbol{n_{side,A}}$ and $-\boldsymbol{n_{side,B}}$ are each vectors that point at a region on the nucleosome side that interacts with the other nucleosome (note that $\boldsymbol{n_{side,B}}$ has to be reflected about the origin for this to be true, hence $-\boldsymbol{n_{side,B}}$ is actually the vector pointing at nucleosome A). In general, the vector that goes from the origin of nucleosome A to nucleosome B ($\boldsymbol{\ell}$ in Figure 3), goes through a particular region on the surface of the nucleosome. If the two nucleosomes have face to face overlap, this vector points to a region of a core histone (H2A, H2B, H3, and H4, excluding their protruding N-terminal tails, see Figure 4) on the top of the nucleosome (core histones marked with "1") or bottom of the nucleosome (core histones marked with "2"). Meanwhile, for side to side overlap, the vector points at a DNA base pair or histone N-terminal tail (we do not consider the C-terminal tails on the other end of the polypeptide and henceforth refer to the N-terminal tails as "tails" for convenience) that sticks out of the side of the nucleosome (see Figure 4). Our goal in this section is to give the regions of interaction (i.e. where $\boldsymbol{\ell}$ points) biological meaning.

A. Using side to side overlap to give tail interactions

Consider the vector $\boldsymbol{n_{side}} = \boldsymbol{n_{side,A}}$ or $-\boldsymbol{n_{side,B}}$ as shown in Figure 3. Since this vector is in the plane of a given nucleosome face, it has nonzero $\boldsymbol{u}$ and $\boldsymbol{v}$ coordinates. Denote these coordinates as *u* and *v* and define an angle $\theta = \arctan\left(\frac{v}{u}\right) \in (-180°, 180°]$ (here we use the four quadrant arctangent). We also define a parameter *n* that denotes the number of helical turns of DNA that correspond to θ. Since there are approximately four turns of DNA from the dyad to the other side of the nucleosome, we let

$$n = \frac{\theta}{45°}$$

So that $n \in (-4,4]$ helical turns. This can help us identify what regions of DNA on the nucleosomes interact with each other. For example, if $n = 1.5$, the vector $\boldsymbol{n_{side}}$ points to the minor groove of the DNA that wraps around the nucleosome. If the minor groove faces inward, the major groove faces outward, and so the major groove of DNA interacts with the other nucleosome.

In addition, either θ or n can be used to determine the nearest histone tail anchor location to the region of interaction. In the nucleosome core particle, each of the four core histone proteins has two copies. Since each has a tail, there are a total of eight histone tails that are anchored somewhere on the nucleosome. We can therefore determine the tail on a nucleosome that interacts with another nucleosome by determining the closest tail to the given θ or n using the

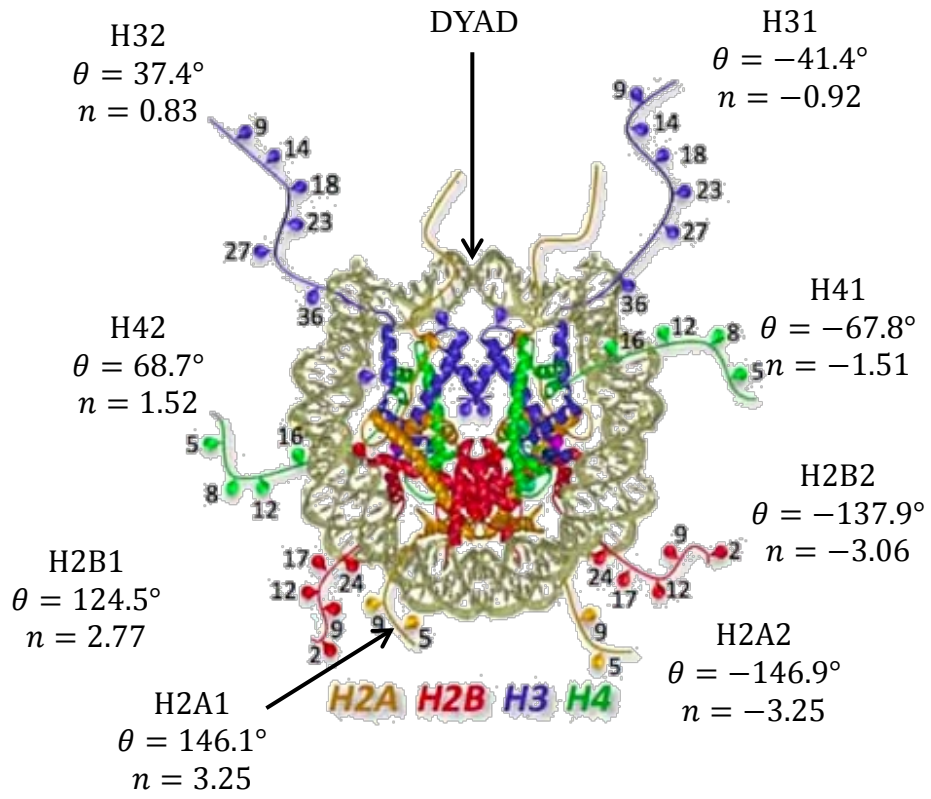

Figure 4: Locations of histone tails with respect to the angle θ and number of helical turns *n* from the dyad. While these coordinates are taken from the PDB reference locations, the tails are very flexible in general and tend to move from where they are anchored, meaning that the locations defined here are not rigid (adapted from [3]).

structure from [1], shown in Figure 4. For example, when $n = 1.5$ turns, the H42 histone tail would be considered the tail of interaction. While this calculation is useful, it is somewhat constrained by the fact that the tails do not point directly outward from where they are anchored and may move around. For example, two nucleosomes interacting at $\theta = 50°$ could be the result of the H42 tail sliding towards the dyad, the H32 tail sliding away from the dyad, or an interaction between the outward facing major grooves of DNA. The only way to tell definitively which case is by looking at the structure.

B. Using face overlap to give histone core interactions

Unlike in side to side interactions, where we only consider the direction of $\boldsymbol{\ell}$ on the face of the nucleosome, the normal direction of $\boldsymbol{\ell}$ plays an important role in mapping face overlap to biological regions. Considering $\hat{\boldsymbol{\ell}}$ as the unit vector pointing from one origin to the other, let $\boldsymbol{p} = \hat{\boldsymbol{\ell}} * \frac{h_{nuc}}{2|\ell_n|}$, where h_{nuc} is the height of the nucleosome (44 Å) and ℓ_n is the normal coordinate of $\boldsymbol{\ell}$ (we are simply rescaling $\boldsymbol{\ell}$ so that it corresponds to a point at the top or bottom of the nucleosome). Here $\boldsymbol{p}$ is the approximate point at which the top or bottom of the nucleosome interacts with the other nucleosome, which we refer to as the point of contact from here forward. Note that if there is a large amount of face overlap, $\boldsymbol{p}$ is but one of many points where interaction occurs. As in the side interaction case, the point $\boldsymbol{p}$ can be matched with the approximate centers of each of the core histones, as well as the "acidic patch," the negatively charged region near the H2A and H2B dimer at which nucleosomes have shown a propensity to

Table 1: Locations of core histones. AP denotes acidic patch between H2A and H2B

Histone	H31	H41	H2A1	H2B1	H32	H42	H2A2	H2B2	AP	AP (–)
$\boldsymbol{u}$	10.1	1.5	–8.1	–17.1	10.4	1.8	–7.7	–16.5	–6.7	–5.9
$\boldsymbol{v}$	–17.1	–16.5	10.9	5.9	17.0	16.5	–10.9	–6.0	7.0	–7.2
$\boldsymbol{n}$	0.9	4.8	15.1	15.0	–2.4	–6.7	–17.2	–17.6	20.3	–22.4

interact [4]. Table 1 shows the mean locations of each of these core histones, computed by taking the average location of all the amino acid C_α atoms that make up the given core histone (i.e. the protein chain excluding the tail).

VI. Using the algorithm – key results

We used the algorithm to study both the amount of interaction between nucleosomes and the region(s) where interaction occurs in crystals, fibers, and MCMC simulations.

X-ray crystal structures were chosen from data sets available through the Protein Data Bank. A central nucleosome was selected, and another nucleosome of the same type was considered part of the crystal lattice if it had an atom within 5 Å of an atom on the central nucleosome. This allowed for a range of 3-13 nucleosomes within a crystal structure. We considered 72 crystal structures, all of which are listed in appendix B.

Fiber structures were adapted from electron distribution data collected from cryogenic electron microscopy (cryo-EM) in [5].

Simulation structures were generated from room temperature Markov Chain Monte Carlo (MCMC) simulations of nucleosome arrays of 13 NCPs with 30 base pair linker DNA. In MCMC simulations, a change in rigid body parameters is proposed and accepted if it decreases the energy of the structure and (with some exceptions) rejected if the change increases the energy (see [6] for more details). This procedure can generate arbitrary numbers of simulated structures, and in our case we analyze interactions over approximately 22,000 simulated structures to get a sense of how nucleosomes are configured on average.

A. Nucleosomes in crystals

Figure 5 shows a characteristic crystal structure. Although this structure (for 1kx5) contains 13 nucleosomes, other structures that have fewer nucleosomes within the 5 Å cut off from the central nucleosome are similar, only with some nucleosomes missing from the non-1kx5 lattice. As shown in Figure 5, the crystal structures in general display two patterns of overlap, shown as

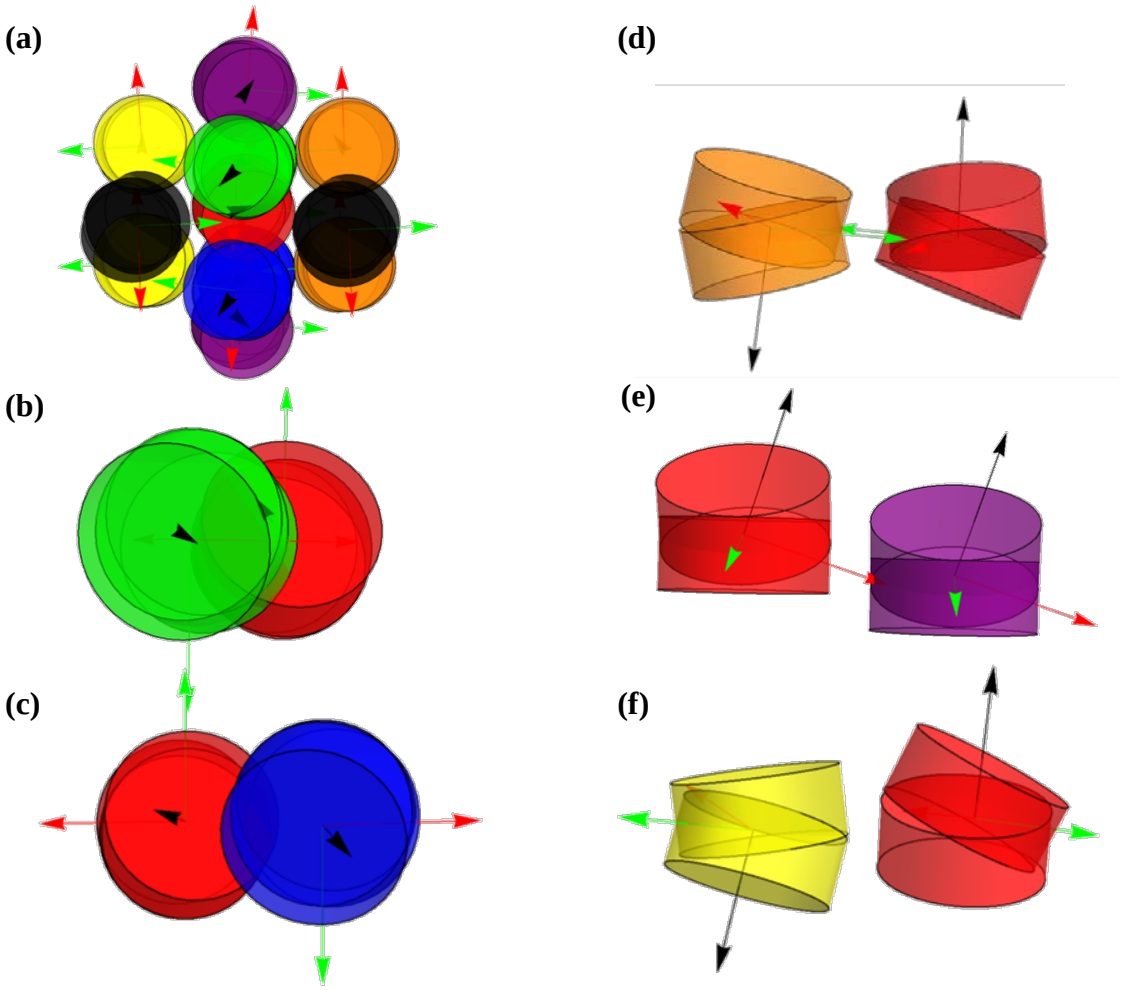

Figure 5: Representative overlap patterns in crystal structures. (a) The 1kx5 crystal structure. Note that two more nucleosomes are hidden behind the central plane containing the red nucleosome. These two are simply the mirror images of the green and blue nucleosomes shown and when counted make for a total of 13 nucleosomes in the crystal. **(b-c)** Face overlap patterns, with **(b)** the highest amount of overlap (about 50%) and **(c)** the second highest (about 15%). **(d-f)** Side overlap patterns, with **(d)** the most amount of side overlap (slightly less than 90%) corresponding to the two nucleosomes immediately connected to the central one by DNA [7] **(e)** the second highest (about 70%) and **(f)** the third highest (about 40%). In all cases, the green axis of the nucleosome points to the dyad.

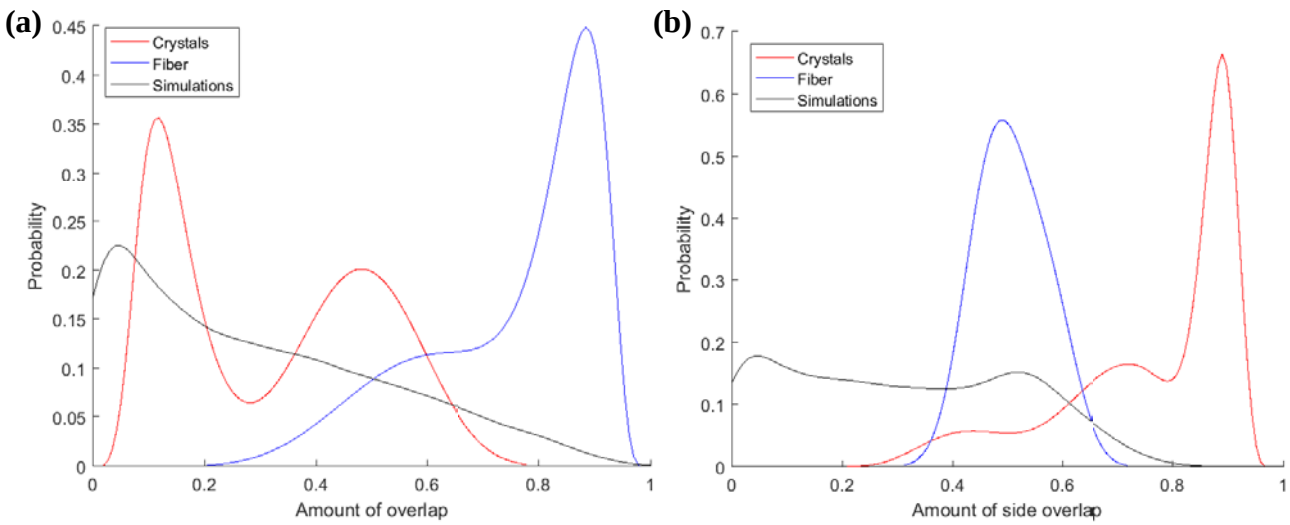

Figure 6: Distributions of (a) overlap and (b) side overlap in crystals (red), fibers (blue), and simulations (black). For crystals, two distinct peaks are observed in overlap, corresponding to the two overlap patterns in Figure 5(b-c), and three peaks are observed for side overlap, which correspond to the three patterns in Figure 5(d-f). In fibers, two peaks of overlap are observed because of the tetrameric structure (although this does not appear in side overlap). Simulations display broad distributions, generally skewed towards lesser overlaps.

green (largest) and blue (second largest). Crystal structures generally show three patterns of side overlap, shown as orange (maximum side overlap), purple (second most side overlap), and lastly yellow (least side overlap). The two patterns of face overlap and three patterns of side, or lateral, overlap are present across all of the crystals and can be seen in the distributions of overlap and side overlaps for crystals (Figure 6, red distributions).

The most obvious natural question here is why these crystals pack in the way they do; that is, whether or not there is any biological relevance to the packing. To answer this, we use the part of our algorithm that maps contact to biological regions. Figure 7 shows the contact points (i.e. the points ***p*** in section V, part B) on the top face of the nucleosome in comparison to the locations of the core histones and acidic patch (i.e. the negatively charged region in the H2A/H2B dimer). Several studies, among them [4,8], have demonstrated that nucleosomes in crystals tend to interact via contacts of the positively charged H4 tail with the acidic patch. Our results in Figure 7 support this, as they show a propensity for nucleosomal interaction near the acidic patch and

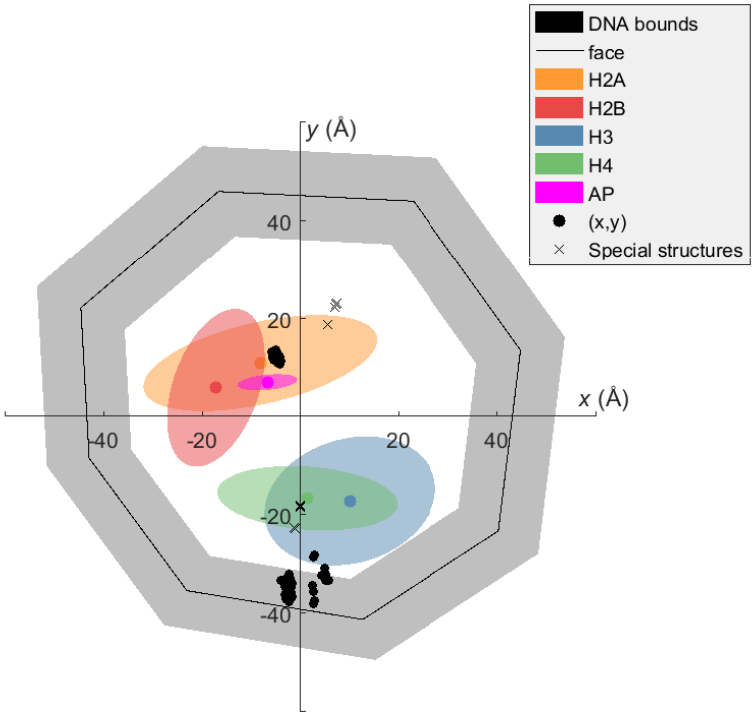

Figure 7: Mapping crystal points of contact on the top face of the nucleosome to biological regions. Observe that contacts are split between the regions in the top and bottom half of the plane. These correspond to the acidic patch, shown in magenta, and H4 histone tail, respectively. Here the H4 histone is shown in green, and contacts on the edge of the nucleosome close the H4 center indicate a high probability that contact is occurring via the H4 tail. Note that some crystal structures, marked with an 'x,' appear to have contacts on the H4 core histone and H2A histone; similar structures are profiled in [8]. Note that from here forward, we dispense with ***u*** and ***v*** notation and refer to the former ***u*** axis as ***x*** and the former ***v*** axis as y.

the H4 tail, which, although mobile, can safely be assumed to lie outward from the H4 core histone. Observe also the correspondence between Figures 5 and 7: the contacts on the acidic patch of the top face (shown in Figure 5(b), where the acidic patch of the central nucleosome in red contacts the H4 tail of the green nucleosome) tend to have higher overlap coefficients than those on the same face with the H4 tail (shown in Figure 5(c)). This means that more of the acidic patch on the upper face of the nucleosome is covered up by another nucleosome than the acidic patch on the lower face.

Although most of the nucleosomes in crystals follow this general pattern of interaction, we found six structures, 3lz0, 3rel, 3lz1, 3rej, 3rek, and 3utb, that had different interaction patterns. As shown in Figure 7, these structures, whose points of interaction are marked with an 'x,' tended to

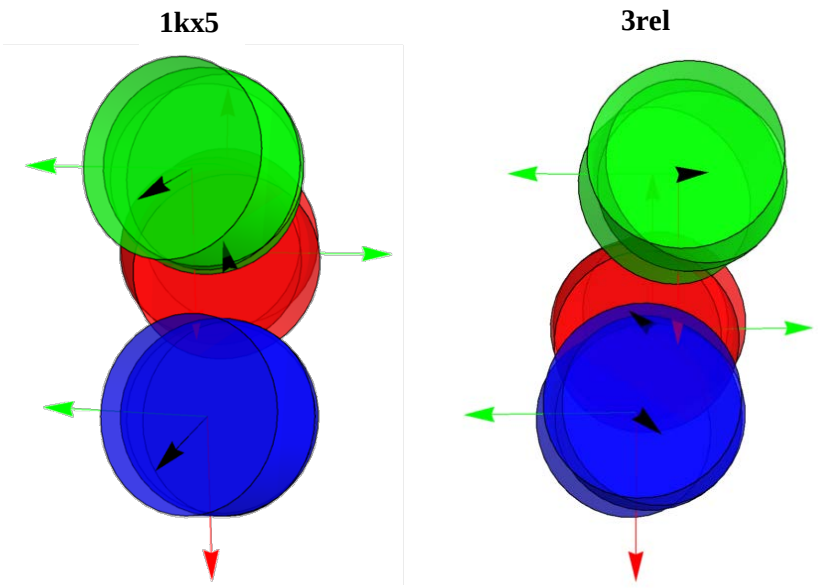

Figure 8: Comparison of 1kx5 with unusual structure 3rel. The 10 Å shift described in [8] is apparent here, with the green nucleosome shifting upward in 3rel so that the acidic patch of the red nucleosome is no longer in contact with the green H4 tail. Meanwhile, the blue nucleosome also shifts upward so that regions on the H4 and H2A core histones pair up individually for contact [8]. In the 3rel structure, the H2A core on the central (red) nucleosome pairs with a region on the H3/H4 core on bottom face of the green nucleosome, and a region of H3/H4 on the central (red) nucleosome pairs with a region of H2A on the bottom face of the blue nucleosome.

interact on the H3/H4 *core histones* (not the tails) and H2A core histone. While none of these structures are specifically discussed in [8], the authors there discovered structural differences due to DNA sequence alterations, stating that a shift caused the acidic H2A-H2B elements and positively charged H3/H4 tail elements to pair up with other elements of the H2A and H2B C-terminal tails [8]. Although we do not account for the C-terminal tails here, we see contact on the *top face* in H3 and H2A, which indicates the H3 on the top face pairs with H2A on the bottom face and vice versa, confirming the results in [8]. We also observe a translational shift upwards from Figure 5, so that now the point of larger overlap in these different structures occurs when H3 on the top (red) face pairs with H2A on the bottom (blue) face of another nucleosome. This shift means that the blue nucleosome has more overlap with the red (central) nucleosome than the green one in the atypical structures, as depicted in Figure 8.

These structures also showed slight differences in their points of contact when considering side overlap, but these minor shifts did not tend to change the overall lateral interactions between the structures. Using our algorithm to map the side overlap direction to a number of DNA turns *n*, we can get an idea of the contact patterns that each of the side overlaps in Figure 5(d-f) correspond to. In Figure 5(d), the contact points occur about one turn off the dyad, and we see from Figure 9 that this is likely an interaction between the H3 tail on one nucleosome with the major groove of the DNA on the other that occurs at $(n_c, n_o) = (+0.8, -0.5)$ or $(-0.5, +0.8)$ (note that if $n = 0.5$, the minor groove is contacting the nucleosome, leaving the major groove to

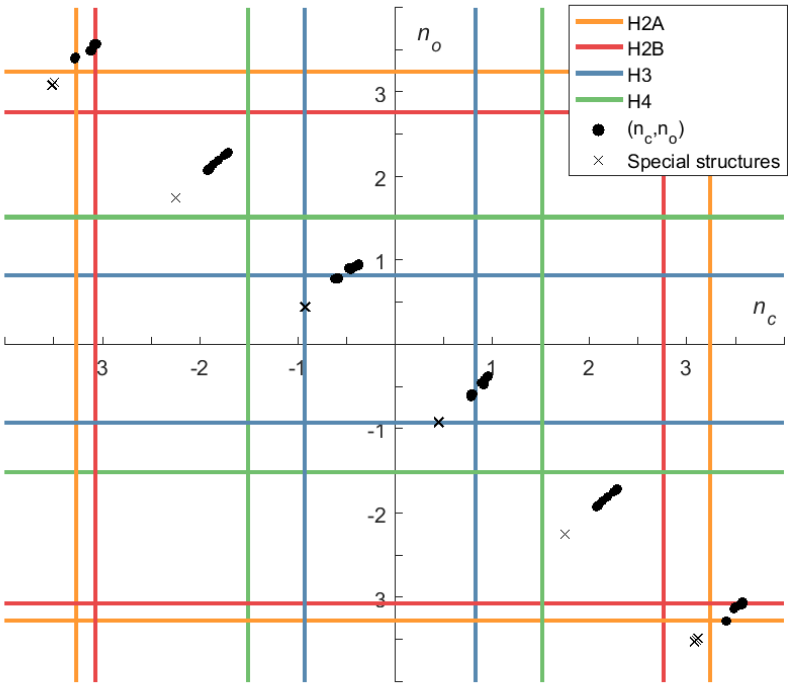

Figure 9: Mapping points of contact on the sides of the nucleosome to the (approximate) tail locations. Each ordered pair (n_c, n_o) denotes the number of turns at which the central nucleosome ('c') interacts with another nucleosome ('o') in the lattice. As expected from Figure 5, there are six distinct clusters since there are six nucleosomes surrounding the central atom (note that the top half of the plane is simply a reflection of the bottom half across the line y=x). The three clusters correspond to the H3 tail interacting with the major groove of the DNA, the minor grooves of the DNA interacting, and the H2A/H2B tails with the major groove of the DNA, although this last interaction is somewhat unclear. As before, some structures are unusual in that they fall below the y=-x line, signaling a slight shift in their interaction patterns. These structures are marked with an 'x.' Note that there are approximately 230 significant side interactions across all of the crystal structures, making for about 30-40 points in each group shown above.

face outwards and interact with the other nucleosome). As discussed in [7], this first type of interaction corresponds to two nucleosomes that are connected directly together with DNA, making it obvious why this type has the highest side overlap.

The second type of interaction, occurring at $(n_c, n_o) = (\pm 2, \mp 2)$ and shown in Figure 5(e) seems to be between the two minor grooves of DNA at turns. Finally, a third type of interaction (from Figure 5(f), around $(n_c, n_o) = (+3.5, -3)$ or $(-3, +3.5)$) might involve the H2A/H2B tails interacting with the major groove of the DNA.

As in face overlap, it can easily be seen (especially for the first two groups, which involve the largest side overlaps) that some structures, marked with an 'x' in Figure 9, fall outside of the normal pattern. In this case, we found that the six structures that had atypical face overlap patterns also had atypical side overlap patterns. These six were joined by 1kx4 and 4z5t (all of which are marked in Appendix B) in having points of interaction below the $y = -x$ line. This means that their interactions are flipped; for example, whereas in a typical structure a nucleosome might interact with the central nucleosome via its H3 tail and the central nucleosome's DNA major groove, a nucleosome at the same position in the lattice in an atypical structure interacts from its DNA major groove to the H3 tail of the central nucleosome.

B. Nucleosomes in fibers

Nucleosome configurations taken from cryogenic electron microscopy (cryo-EM) show different behavior from those captured in crystals. As shown in Figure 10, a fiber of 12 nucleosomes with 40 base pairs of linker DNA arranges itself into two distinct stacks (colored red and blue in the figure). The two distinct stacks explain why the side overlap is more or less normally distributed around 50% (Figure 6(b, blue distribution)), since the side overlap being measured between nucleosome *i* and *i*+1 consistently measures the amount of side overlap from one side of the fiber

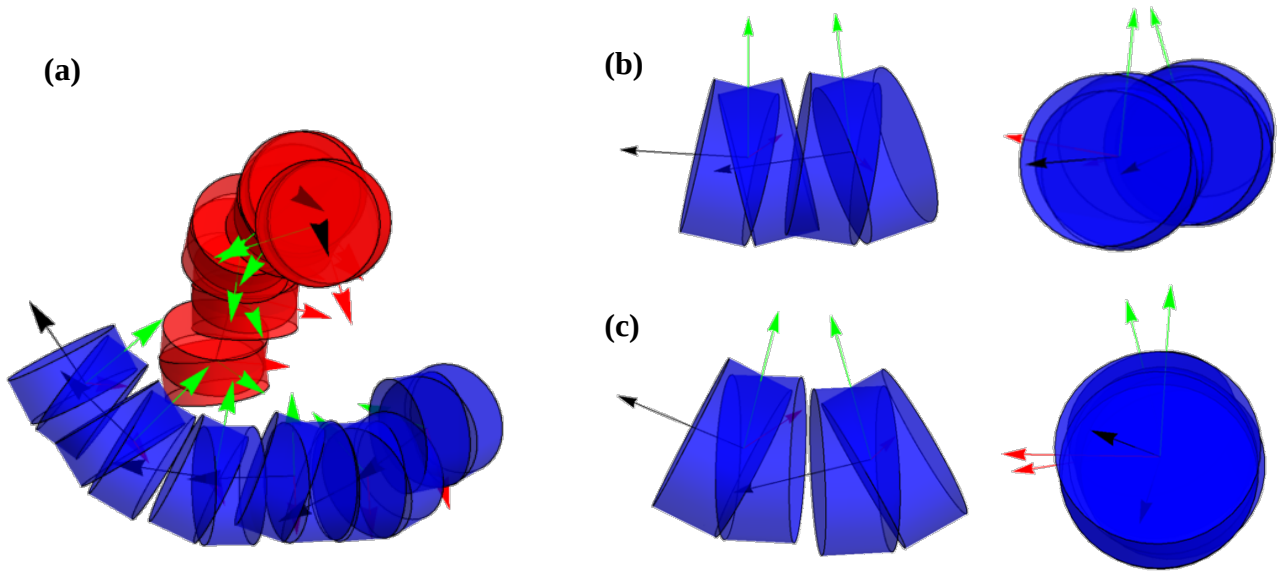

Figure 10: Nucleosome configurations and face to face overlaps in a cryo-EM structure. (a) The entire structure is shown, with the alternating nature made clear by the use of red for odd numbered nucleosomes and blue for even numbered ones. The two face to face overlap patterns are shown in **(b)**, which has the most overlap and occurs between nucleosomes in the same tetrameric unit, and **(c)** which has less overlap and is characteristic of nucleosomes interacting between tetrameric units. Note that in **(c)**, the wedge shape can be deceiving. Looking at the normal vectors in this case makes it clear why this configuration has less overlap when the nucleosomes are modeled as cylinders rather than wedges. The two patterns of overlap are evidence of the tetrameric structure of the fiber [5].

to the other. Note also that this side overlap tends to occur around $(n_i, n_{i+1}) = (0.3,0)$, where the notation now defines the number of helical turns on the DNA surrounding nucleosomes *i* and *i*+1, respectively, where the side interaction occurs. Face overlap coefficients reveal that each of the stacks is itself an alternating series, with one set of nucleosomes (Figure 10(c)) having a smaller amount of overlap (about 60%) and the next pair having a higher amount (about 90%, (Figure 10(b))). This pattern explains the bimodal distribution of overlaps shown in Figure 6(a, blue distribution). Both amounts of face to face overlap are significantly higher than in the crystal structures, where magnitudes of 10% and 50% were common. This might owe to the inclusion of a fifth histone, H1, in the fiber, as discussed in [5].

In [5], the authors found different biological regions making up the two types of interaction, with interactions within a tetrameric unit coming between the H2A/H2B regions and interactions

across units coming between the acidic patch and H4 tail. Our analysis partially confirming this is shown in Figure 11, where we map the points of interaction between nucleosomes onto the nucleosome face. We found that nucleosomes interacting within the same tetrameric unit (labeled with black points) tended to interact near their origins and H2A/H2B dimers, as stated in [5]. However, our method did not find evidence of an acidic patch to H4 tail interaction between tetrameric units, instead mapping these interactions to the H3/H4 region on both nucleosomes (red points). This, however, does not mean that the interaction did not occur. As Figure 10(c) shows, nucleosomes interacting across fibers have high face to face overlap, and thus mapping that overlap to a single point risks over-simplifying the area where the two nucleosomes are in contact. Furthermore, treating the nucleosome as a cylinder instead of a wedge also introduces some error, as Figure 10(c) shows high overlap between the *wedge* faces, but looking at the

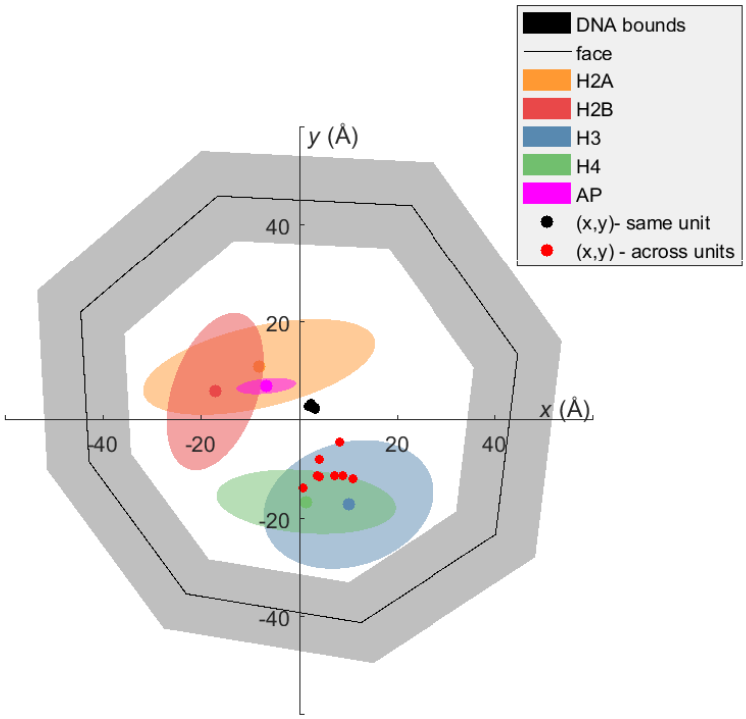

Figure 11: Overlap patterns in cryo-EM structures. As discussed in [5], the regions of overlap are quite different when nucleosomes interact within the same tetrameric unit (black points) than when they interact across different units (red points). The black points show interaction within tetrameric units tends to happen near the origin and H2A/H2B dimer, as discussed in [5]. Meanwhile, we found interaction across different units tends to happen near the H3/H4 interface.

misalignment of the normal vectors demonstrates that there is less overlap between the two when they are considered as cylinders.

In sum, our analysis for fibers supports some of the conclusions reached in [5], namely that the cryo-EM structure can be broken up into tetrameric units and that interactions between nucleosomes that fall within the same unit tend to happen near the H2A/H2B dimer of both nucleosomes. We were unable to capture the H4 tail to acidic patch interaction that was found between distinct tetrameric units, however, likely owing to our oversimplification of a large region of overlap into a single point of contact and treatment of nucleosomes as single cylinders instead of pairs of H2A-H2B-H3-H4 cylinders.

C. Nucleosomes in simulated configurations

As shown in Figure 6 (black distributions), simulated structures behave differently than X-ray and cryo-EM structures in that the distribution of overlaps and side overlaps for simulations tends to be broad and skewed towards 0. This occurs because the structures are simulated at room temperature, have a larger range of motion, and have on average about twice as much space between overlapping nucleosomes than cryo-EM and X-ray structures. Given these factors, it is remarkable that the room temperature simulations capture the same biological interaction patterns as the X-ray and cryo-EM structures. As shown in Figure 12(a), simulated fibers with an array of 13 nucleosomes and 30 base pairs of DNA linkers tend to arrange themselves in three stacks, with the most overlap occurring between nucleosome i and i+3. Furthermore, the points of overlap between the i^{th} and $(i+3)^{th}$ nucleosome generally map to regions between the H2A/H2B dimer (acidic patch) and the H4 core histone (Figure 12(c)). These face to face interaction regions are similar to those for cryo-EM structures, where the large density of interaction near and below the origin in the simulations matches the cryo-EM overlap that occurs

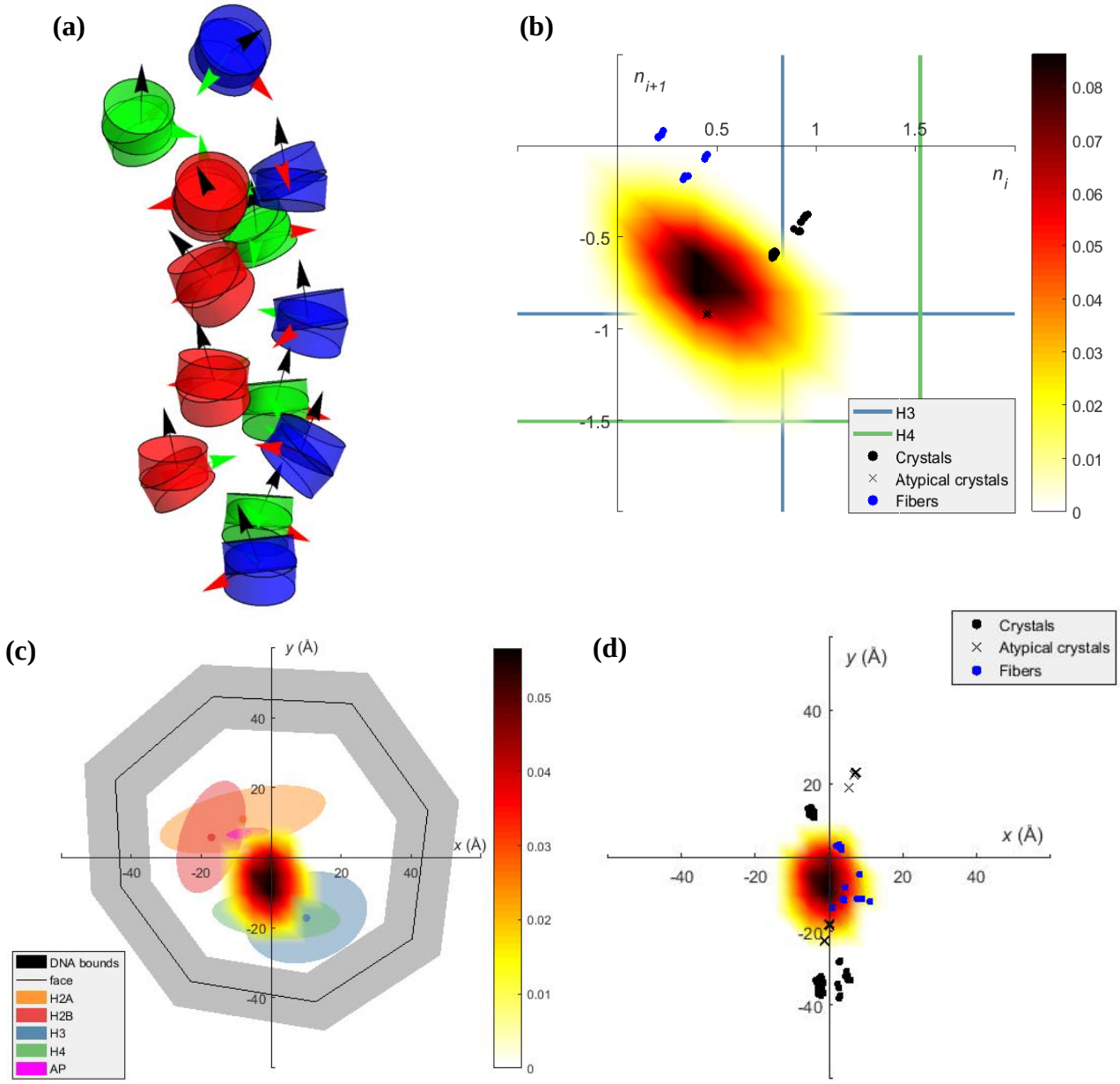

Figure 12: Simulated structures for 30 base pair DNA linkers. (a) A representative structure (with face to face overlap values near the average values over all of the simulations) shows the 13 NCP fiber forming three stacks, colored for convenience. The largest overlap patterns are therefore between nucleosome *i* and *i*+3, and the largest side overlap occurs between nucleosome *i* and *i*+1. (**b)** A heat map of biological regions of side overlap shows that the interaction of H3 tails and DNA is key to understanding side interactions. Overlaid plots of crystal structure and cryo-EM interactions show the simulations to be in best agreement with the crystals. **(c)** A heat map of biological regions of overlap throughout the simulations on the top face of nucleosomes shows that overlap generally occurs in the region between the acidic patch on the H2A/H2B dimer and the H4 core histone. **(d)** Comparing the overlap regions on the nucleosome face to those of crystal structures and cryo-EM fibers shows that the simulations realistically model cryo-EM face to face overlap.

both within and between tetrameric units (see Figure 12(d) for a side by side comparison). The

similarity is not surprising given the similarity of the structures themselves; the representative configuration from the simulations in Figure 12(a) shows a face to face stacking pattern that is more similar to the cryo-EM structure in Figure 10(a), where one nucleosome in a stack is directly beneath the prior one, than to the crystal structure in Figure 5(a), where two nucleosomes compete to both overlap the central nucleosome on its top face.

The similarity between the MCMC simulated fibers and the X-ray structures is obvious when the side overlap interaction patterns are examined and can also be explained by comparing the structures. As shown in Figure 12(b), side overlaps in simulations tend to occur between the major groove of the DNA in nucleosome *i* (i.e. $n_i = 0.5$, the minor groove points in and the major groove points towards nucleosome *i*+1) and the H3 tail of nucleosome *i*+1 ($n_{i+1} = 0.8$). In the simulations, there is only one cluster of side overlap densities because we only consider the *i* to *i*+1 nucleosome side overlap and only move in one direction on the fiber (the direction of the DNA sequence, from blue to red to green in Figure 12(a)). Note the similarity with X-ray crystal structure side interactions, shown in Figure 5(d), in which nucleosomes in the lattice that had the largest side overlap also interacted via the H3 tail and DNA major groove. This is not surprising since in both the X-ray and simulation structures, the nucleosomes that have large side overlap are immediately connected by DNA, as shown for X-ray structures in [7].

In sum, despite the limitations of MCMC room temperature simulations, the structures they generate are able to capture the overlap patterns that occur between the faces and sides of adjacent nucleosomes across both cryo-EM fibers and X-ray crystals.

VII. Conclusion

We began by modeling the nucleosome as a polygonal cylinder and defined its face and side by an octagon and rectangle, respectively. This was key in computing the relative amount of overlap

and side overlap between two nucleosomes. We next defined a vector that allowed us to compute an approximate point where the nucleosomes are in contact both on their faces and sides and map it to biological regions.

Information about the direction of interaction in turn allowed us to analyze X-ray crystal structures, cryo-EM fibers, and MCMC simulated structures. X-ray structures were found to exhibit two kinds of face to face overlap and three kinds of side overlap, with the face to face overlap occurring via the H4 tail and the acidic patch in most cases and the strongest side to side overlaps occurring when DNA connected the two nucleosomes. Several unusual X-ray structures, characterized by H2A-H3/H4 core histone interactions, were also discovered.

Cryo-EM structures were confirmed to exist as tetrameric units of four nucleosomes which interacted differently within the same unit than between different units, and the side to side overlap in cryo-EM structures proved to take on one value as the fiber oscillated back and forth between two stacks.

Simulations, which generated broad distributions of overlaps and side overlaps, were able to replicate biological interaction regions from both types of measurements, as face overlaps were most densely accumulated in simulations at the same points that they occurred in cryo-EM, whereas side overlaps exhibited the same behavior in simulations as they did in X-ray crystals.

VIII. Acknowledgements

I want to thank Dr. Olson and Stefjord Todolli for all of their help this summer. I also want to thank DIMACS, Dr. Lazaros Gallos, and Parker Hund for organizing the REU. Thank you to the NSF (grant CCF-1559855) and USPHS (grant GM 34809) for funding.

Appendix A – Computing the mid-frame (adapted from [2])

Given nucleosomes A and B with orthonormal basis vectors $\boldsymbol{u_i}, \boldsymbol{v_i}$, and $\boldsymbol{n_i}$ and origins $\boldsymbol{o_i}$, the first step in computing the mid-frame is to compute the angle between the two normal vectors.

$$\Gamma = \arccos(\boldsymbol{n_A} \cdot \boldsymbol{n_B}) \tag{A-1}$$

With the condition that the sign of the angle is positive if the cross product of the two normals points in the positive $\boldsymbol{z}$ direction and negative if it points in the negative $\boldsymbol{z}$ direction (i.e. Γ and $(\boldsymbol{n_A} \times \boldsymbol{n_B}) \cdot \boldsymbol{e_z}$ have the same sign, where $\boldsymbol{e_z} = (0\ 0\ 1)^{\mathrm{T}}$ is the unit vector in the $\boldsymbol{z}$ direction). Note that this convention is arbitrary and is used to determine the orientations of the nucleosomes with respect to each other.

Next the so-called "roll-tilt" (RT) axis is defined:

$$\boldsymbol{rt} = \boldsymbol{n_A} \times \boldsymbol{n_B} \tag{A-2}$$

After normalizing the RT axis, the reference frame of nucleosome A is rotated by $\Gamma/2$ about the RT axis and B by $-\Gamma/2$ about the RT axis so that their $\boldsymbol{n}$ axes are now aligned. Denote the new coordinate frames generated by rotation $\boldsymbol{u_i'}, \boldsymbol{v_i'}$, and $\boldsymbol{n_i'}$ (note that $\boldsymbol{n_A'} = \boldsymbol{n_B'} = \boldsymbol{n_m}$, the mid-frame normal axis). Figure A1 depicts the alignment of the frames at this point in the process.

Next the angle between the transformed $\boldsymbol{v}$ axes is computed:

$$\Omega = \arccos(\boldsymbol{v_A'} \cdot \boldsymbol{v_B'}), \tag{A-3}$$

where the sign of Ω is the same as the sign of $(\boldsymbol{v_A'} \times \boldsymbol{v_B'}) \cdot \boldsymbol{n_m}$. That is, the sign of Ω is the sign of the direction of the cross product of the two $\boldsymbol{v'}$ vectors with respect to the mid-frame normal. In order to find the mid-frame $\boldsymbol{v}$ axis, we simply rotate $\boldsymbol{v_A'}$ by $\Omega/2$ (or $\boldsymbol{v_B'}$ by $-\Omega/2$) about the mid-frame normal axis. The rotated $\boldsymbol{v}$ axis is the mid-frame $\boldsymbol{v}$ axis,$\boldsymbol{v_m}$.

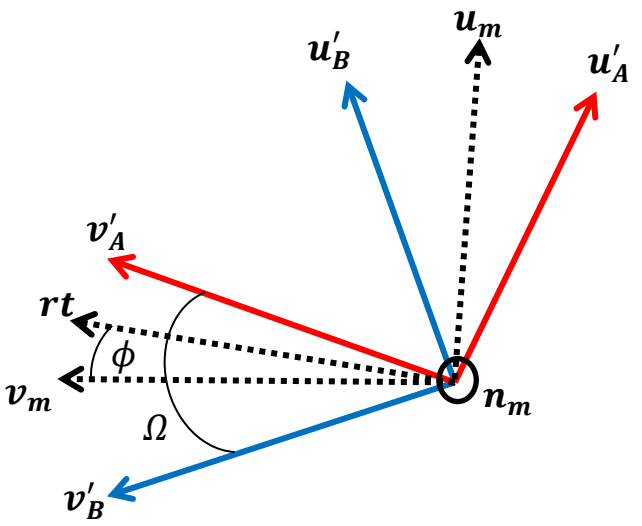

Figure A1: The alignment of the axes after rotating both coordinate systems about the RT axis to obtain a common normal (adapted from [2]).

The mid-frame $\boldsymbol{u}$ axis can be found either by rotating $\boldsymbol{u'_A}$ by $\Omega/2$ (or $\boldsymbol{u'_B}$ by $-\Omega/2$) about the mid-frame normal axis (i.e. by following the same procedure used to find $\boldsymbol{v_m}$) or by using the RT axis. Suppose the RT axis is separated from $\boldsymbol{v_m}$ by an angle ϕ. Then,

$$\phi = \arccos(\boldsymbol{rt} \cdot \boldsymbol{v_m}) \tag{A-4}$$

Where the sign of ϕ is the same as the sign of $(\boldsymbol{rt} \times \boldsymbol{v_m}) \cdot \boldsymbol{n_m}$. That is, the sign of ϕ is the sign of the direction of the cross product of the $\boldsymbol{rt}$ and $\boldsymbol{v_m}$vectors with respect to the mid-frame normal. Now, by this formulation, the $\boldsymbol{u_m}$ axis is $\frac{3\pi}{2} + \phi$ radians from the RT axis. Rotating the RT axis by $\frac{3\pi}{2} + \phi$ radians about $\boldsymbol{n_m}$ gives the mid-frame $\boldsymbol{u}$ axis, $\boldsymbol{u_m}$.

This mid-frame origin is much easier to compute. It is simply the geometric center of the two origins:

$$\boldsymbol{o_m} = \frac{(\boldsymbol{o_A} + \boldsymbol{o_B})}{2} \tag{A-5}$$

We have therefore computed a common frame and origin to project the polygon of each nucleosome onto.

Appendix B – List of Crystal Structures

The list below shows the PDB ID of the structure and the number of nucleosomes in its crystal lattice. Structures marked with an asterisk (*) showed deviation from normal interaction behavior as explained in the main text and [8].

1. 1eqz (7)
2. 1f66 (9)
3. 1kx3 (9)
4. 1kx4* (5)
5. 1kx5 (13)
6. 1m18 (7)
7. 1m19 (7)
8. 1m1a (9)
9. 1p34 (5)
10. 1p3g (7)
11. 1p3i (7)
12. 1p3l (5)
13. 1p3o (5)
14. 1p3p (5)
15. 1s32 (5)
16. 2cv5 (13)
17. 2nqb (5)
18. 2nzd (7)
19. 2pyo (9)
20. 3a6n (9)
21. 3afa (9)
22. 3av1 (9)
23. 3av2 (9)
24. 3azf (7)
25. 3azg (9)
26. 3azi (11)
27. 3azl (9)
28. 3c1b (7)
29. 3lja (7)
30. 3lz0* (9)
31. 3lz1* (9)
32. 3mgp (7)
33. 3mgq (7)
34. 3mgr (9)
35. 3mnn (7)
36. 3reh (7)
37. 3rei (7)
38. 3rej* (11)
39. 3rek* (11)
40. 3rel* (11)
41. 3ut9 (11)
42. 3uta (7)
43. 3utb* (11)
44. 3wkj (9)
45. 3wtp (13)
46. 4j8u (7)
47. 4j8v (9)
48. 4j8w (9)
49. 4kgc (7)
50. 4wu8 (7)
51. 4wu9 (7)
52. 4xzq (7)
53. 4z5t* (5)
54. 4z66 (7)
55. 5av5 (7)
56. 5av6 (7)
57. 5av8 (7)
58. 5av9 (7)
59. 5avb (7)
60. 5avc (5)
61. 5b0y (13)
62. 5b0z (13)
63. 5b1l (13)
64. 5b1m (13)
65. 5b2j (7)
66. 5b31 (9)
67. 5b32 (13)
68. 5cp6 (7)
69. 5dnn (7)
70. 5f99 (7)
71. 5jrg (3)
72. 5x7x (13)

References

[1] CA Davey, D Sargent, K Luger, AW Maeder, T Richmond, “Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9 A Resolution,” *J. Mol. Biol* vol. 319, pp. 1097-1113, 2002.

[2] X Lu, MA El Hassan, CA Hunter, “Structure and Conformation of Helical Nucleic Acids: Analysis Program (SCHNAap),” *J. Mol. Biol* vol. 273, pp. 668-680, 1997.

[3] K Luger, “Nucleosomes: Structure and Function,” eLS, 2001.

[4] K Luger, A Maeder, R Richmond, D Sargent, T Richmond, “Crystal structure of the nucleosome core particle at 2.8 Å resolution,” *Nature*, vol. 389, pp. 251-260, Sep 1997.

[5] F Song, P Chen, D Sun, M Wang, L Dong, D Liang, RM Xu, P Zhu, G Li, “Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units,” *Science* vol. 344, pp. 376-80, Apr 2014.

[6] N Clauvelin, P Lo, O I Kulaeva, E V Nizovtseva, J Diaz-Montes, J Zola, M Parashar, V M Studitsky, W K Olson, “Nucleosome positioning and composition modulate *in silico* chromatin flexibility,” *J. Phys. Cond. Mat.* vol. 27, no. 6, Jan 2015.

[7] JM Harp, BL Hanson, DE Timm, GJ Bunick, “Asymmetries in the nucleosome core particle at 2.5 Å resolution,” *Acta Crystallographica Section D*, vol. 56, pp. 1513-34, 2000.

[8] B Wu, K Mohideen, D Vasudevan, CA Davey, “Structural Insight into the Sequence and Dependence of Nucleosome Positioning,” *Structure* vol. 18, pp. 528-37, Apr 2010.