Big Data Analysis on How Protein-DNA Interactions Impact Nucleosome Folding

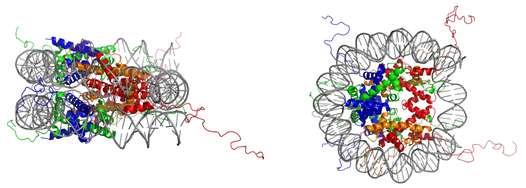
Antonio Bu Sha[1], Advisor: Wilma Olson PhD[2]

[1]Mechanical and Aerospace Engineering, [2]SAS - Chemistry & Chemical Biology, Rutgers University, Piscataway, New Jersey 08854

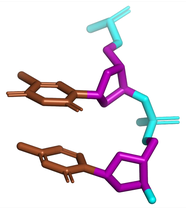
RUTGERS
Aresty Research Center for Undergraduates

Abstract

While it is well understood that the DNA sequence (ATCG) contains the information for how to construct the proteins and RNA that allow most life to function, the role of DNA folding in gene expression is not well understood. DNA is stored in the nucleus by being wrapped around collections of proteins called histones, to form a nucleosome. Our research group is investigating how the specific way that DNA wraps around histones impacts the detachment and subsequent processing of the interacting DNA. Our research group used programs to produce computational models of nucleosomes from the Protein Data Bank (PDB) by analyzing the properties of individual interactions between the DNA and histones. By cataloging and analyzing these interactions, trends across multiple structures can be observed. Preliminary data suggest that DNA often binds asymmetrically to the histone core and may unravel more easily from the side with fewer interactions with histones. This research shows that analyzing histone-DNA interactions will help us better understand DNA folding in nucleosomes. Further research is planned to analyze the effects of non-histone proteins on the way that DNA wraps and unwraps around the histone core.

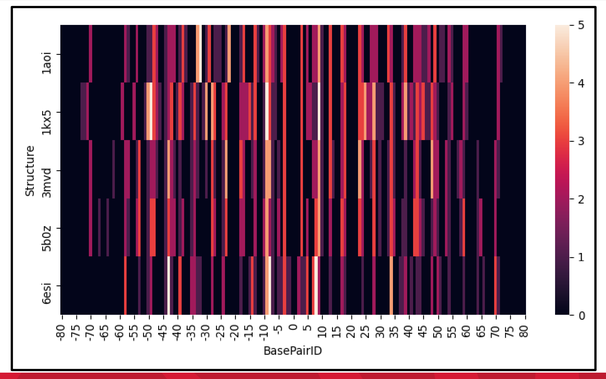


[Figure 1] The images above depict the nucleosome 1kx5, from different viewpoints. The different histones in 1kx5 are color-coded. Chain H3 is red, chain H4 is orange, chain H2A is green, and chain H2B is blue. The DNA is colored grey in these images. [5]



[Figure 2] To the left is an image of two adjacent residues (called nucleotides) from a single strand of DNA. It is color coded, where phosphate groups (PO4) are cyan (blue-green), bases are brown, and sugars are purple. [5]

[Figure 3] The heatmap below shows the frequency of interactions between the DNA and histones of a series of nucleosomes. The x-axis displays the residue number of the interaction, which is the relative location of the interaction in the DNA, in terms of the distance the interaction is from the dyad, or center, of the nucleosome. The y-axis indicates the interactions of different structures, where the brightness of the section on the heatmap corresponds to the frequency of interactions at a specific residue ID on a structure.



Background

- The nucleosome can be described as the transitional structure between unfolded DNA helices and the folded form that DNA takes as the chromosome. When DNA folds, it first wraps around proteins called histones, forming nucleosomes. Strands of nucleosomes are the basic building blocks of chromatin, which is what makes up chromosomes. It is important to know that even though all nucleosomes appear similar, there are many different nucleosomes with subtly different structures. [1]
- Although most DNA research is focused on analyzing the order of bases to study gene sequences, less research focuses on how the physical structure of DNA in nucleosomes impacts gene expression. In order to understand genetics, researchers must better understand the packaging of DNA in chromatin. [2]
- The goal of this research group is to study the interactions between DNA and histones to better understand how DNA folds onto and off a histone core. This work may provide a better understanding of how interactions between DNA and histones impact gene expression.

Figure 4a

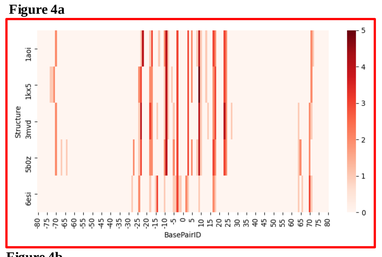


Figure 4b

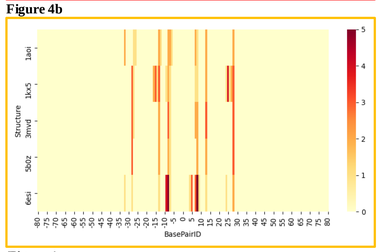


Figure 4c

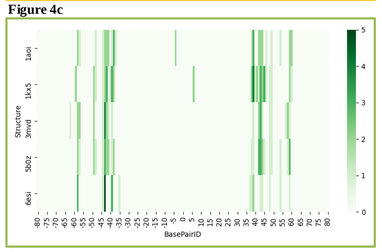
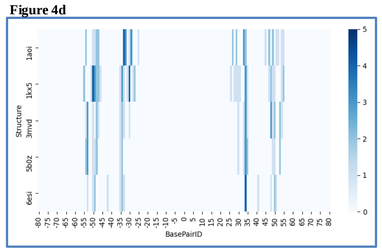


Figure 4d



[Figures 4a, 4b, 4c, 4d] The four heatmaps to the left show the frequency of interactions on specific chains in the histone.
From top to bottom, listed are the interactions at chain: H3, H4, H2A and H2B.
Note the colors of the heatmaps match the respective colors of the chains in Figure 1.

Methods and Materials

Resources Used:

- The Protein Data Bank (PDB) is an online resource that catalogues the 3D shape of proteins, and related organic structures. We used the PDB in order to construct databases of nucleosomes and other structures. [3]
- SNAP, a piece of software from 3DNA, was used to extract numerical information that quantified the interactions between DNA and proteins in nucleosomes from PDB files. [4]
- Code was written in Python to analyze the data that was extracted from PDB files.

Structure	NucleotideAtom	BasePairID	N_Chain	N_BaseName	N_ResidueID	AminoAcidAtom
1aoiA	OP2	-54	I	DA	-54	OG
1aoiA	O5'	-54	I	DA	-54	N
1aoiA	OP1	-43	I	DA	-43	NH1
1aoiA	OP2	-43	I	DA	-43	NH1
1aoiA	OP1	-42	I	DG	-42	N

[Table 1] An example section of the data table built from the information extracted by SNAP from the PDB files. Each property of each reaction was separated into a different column.

Programming Methods:

- After the PDB files for individual structures were converted for use with SNAP, the resulting data were imported into Python. With a data frame, each documented interaction was separated into its composite characteristics, such as the type of protein and DNA involved in the interaction, the location of each interaction along the DNA strand, and more.
- With the data saved in a chart, subsets of data were analyzed by screening out entries that fit or failed to fit the desired set of criteria.
- Python programs were designed and written to be used by people who were not programming experts. This was accomplished with a modular design of the programs, allowing users to select the most ideal functions for the analysis of their data.

Results

- Repeating patterns are visible when viewing Figure 3, with similar areas of high and low interaction frequency, repeating approximately every 10 base pairs. While the structures are not exactly symmetrical, the data show similar patterns of interactions on both sides of the middle base pair, identified as base pair “0”. These patterns are clearer when the data are separated depending on the interacting chain (visible in Figures 4a, 4b, 4c, and 4d).
- Moreover, the frequency of interactions along the nucleosomes correlates with the relative strength of the bond between the DNA and histones, along with the physical distance between the two. For example, the nucleosome “6esi” has a lower interaction frequency relative to the average for other nucleosomes at both ends of its DNA strand, and as such, the DNA is further away from the histone than most other nucleosomes.
- The nucleosome “6esi”, in which the DNA is less attached to the histone core, shows an increase in interaction frequency in phosphate groups in chain H4, while interaction frequency decreases across the rest of the structure, which is visible in the bottom row of Figure 3, 4b, and 5a.

Figure 5a

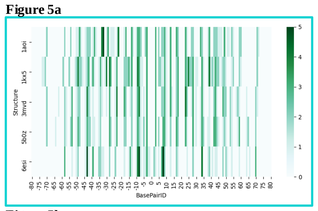


Figure 5b

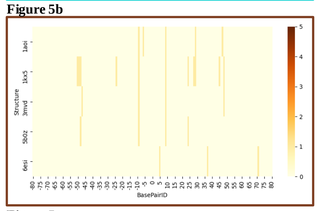
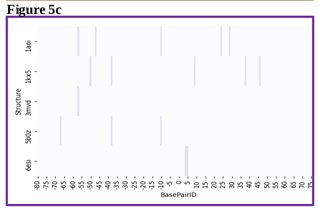
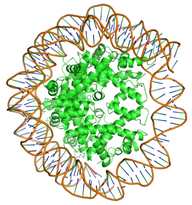


Figure 5c



[Figures 5a, 5b, 5c] The three heatmaps to the left show the frequency of interactions with specific chemical groups in DNA (sugar, base, phosphate). From top to bottom, the heatmaps describe phosphate (PO4), base, and sugar interactions.
Note the colors of the heatmaps match the respective colors of the DNA chemical groups in Figure 2.



[Figure 6] Above is an image of the nucleosome 6esi. While difficult to see in an image, the DNA at the ends of the strand (upper right and lower left) is further from the histones compared to the nucleosome 1kx5. [5]

Conclusion

These data suggest that interaction frequency is closely related to the physical structure of a nucleosome. In addition, the data also suggest that physical changes in nucleosome structure cause reactive changes in other parts of the nucleosome.

Acknowledgements

I would like to thank Prof. Olson and her research group, Stefjord Todolli, Robert Young, Jiyoun Ackerman, for assisting me in my experience. I would also like to thank the Aresty Summer Science Program and Rutgers University for supporting this research project.

References

[1] Annunziato, A. (2008) DNA Packaging: Nucleosomes and Chromatin. Nature Education 1(1):26
[2] Ball, P. (2008, April 28). Pulling our strings. Retrieved July 18, 2020, from https://www.chemistryworld.com/features/pulling-our-strings/3004744.article
[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.
[4] Xiang-Jun Lu & Wilma K. Olson (2003). ‘3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures’, Nucleic Acids Res. 31(17), 5108-21.
[5] The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

RUTGERS