参赛队员姓名： 刘宇慧

中学： 北京市十一学校

省份： 北京

国家/地区：北京

指导教师姓名： 周政 窦向梅

指导教师单位： 中国科学院生物物理所
北京市十一学校

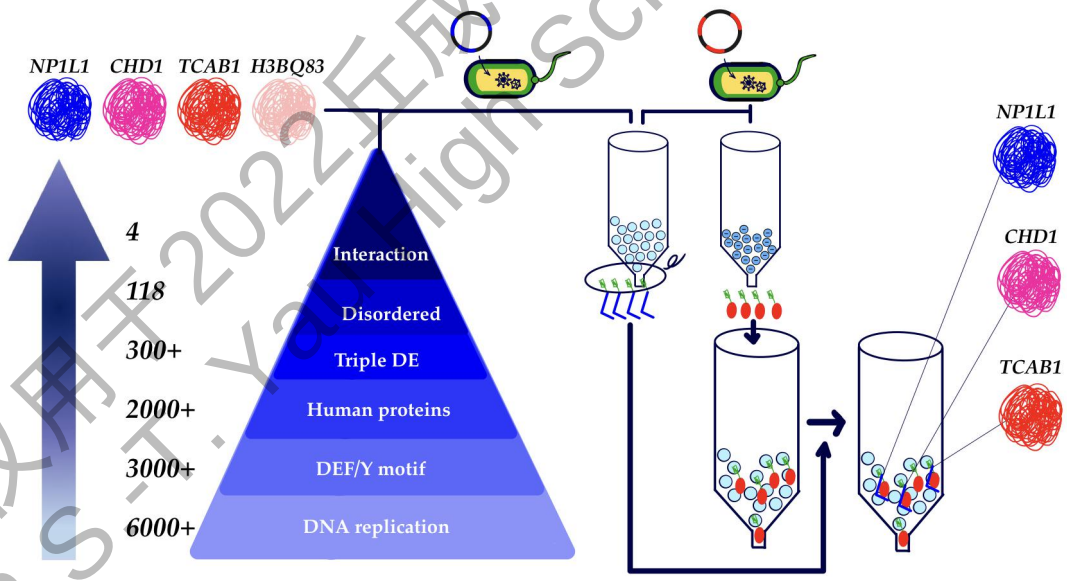
论文题目：A simple, effective approach to screen and identify H2A-H2B binding proteins

A simple, effective approach

to screen and identify H2A-H2B binding proteins.

Abstract

Histone binding proteins regulate histone metabolism and play important roles in multiple biological processes that are linked to human diseases. Here we developed an in silico approach to screen and identified human histone H2A-H2B binding proteins. Target proteins are selected from the protein database and screened based on protein characteristics, which include specific histone binding motifs, protein folding tendency, and overall charges. Then, the binding between H2A-H2B and selected proteins is predicted by AlphaFold2, a computational protein structure prediction method, and further validated by pull-down experiments. Using this approach, we screened the DNA replication-associated proteins and demonstrated that NP1L1, CHD1, and H3BQ83 are H2A-H2B binding proteins. Collectively, this study reports a simple yet effective approach to identify novel H2A-H2B binding proteins.



Keywords: Binding proteins, computer analysis, screening, identification, histones, binding affinity, pull-down experiment, DNA replication, H2A-H2B.

Contents

1. Introduction

A. Overview of H2A-H2B binding protein

Nucleosomes are the basic unit of chromatin, which is the material basis of epigenetic control. Each nucleosome comprises a central histone octamer and a DNA strand surrounding that length in ~145 base pairs. The histone octamer contains two H2A-H2B heterodimers and one $(H3\text{-}H4)_2$ heterotetramer (Luger et al., 1997). (Figure 1A) Because of the condensed structure of histones and DNA (the structure of nucleosomes), the transcription, replication, repair, and other processing of DNA require the reassembling of nucleosomes. These modifications require specific complexes, which can be cataloged into histone chaperone and ATP-dependent chromatin remodeling complexes. Both contain particular recognizing binding regions inferred as binding proteins in this passage. In addition, the binding proteins of the H2A-H2B histone heterodimer will be called H2A-H2B binding proteins.

The H2A-H2B heterodimer contains a DNA binding region (DBR) and a histone binding region (HBR), interacting with the surrounding DNA strand and the H3-H4 histone, respectively. During DNA processing or epigenetic regulations, part of the complex will replace the DNA's position and interact with the DBR to expose the DNA (Figure 1B). Moreover, other parts of the complex may play a role in maintaining the original shape of the nucleosome by reinforcing the binding at the HBR or reassembling the nucleosome and maintaining the spatial arrangement of the genetic information after the process is finished (Huang et al., 2020). The H2A-H2B binding protein will primarily function in the binding along the DBR around the histone.

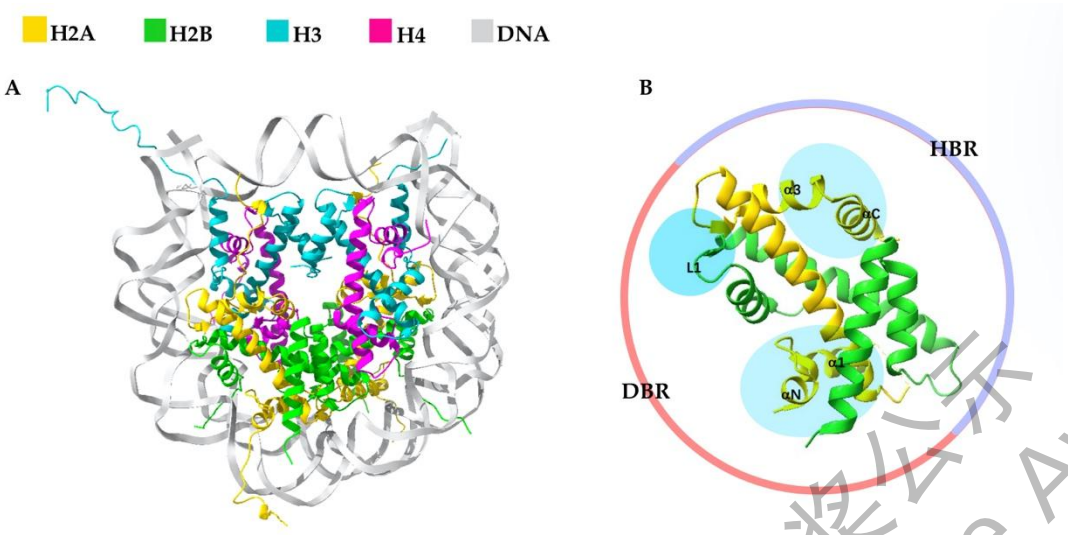


Figure 1 The structure of nucleosome and H2A-H2B dimer.

(**A**) nucleosome structure. The nucleosome structure is modified from PDB 1AO1 and the H2A-H2B structure. Histone and DNA are shown in cartoon and ribbon mode, respectively, and colored as illustrated. (**B**) H2A-H2B structure. The histone binding region (HBR, purple), DNA binding region (DBR, red), and corresponding secondary structure elements are labeled as indicated. The binding sites are marked in blue circles. The darker color indicates a more common binding.

B. Current research about H2A-H2B binding protein

Recent studies have found different nucleosome remodeling proteins and their binding protein, including the Chz1, ANP32e, YL1, and Swc5. The proteins play roles in transcription, DNA replication, and other chromosome remodeling processes.

Chz1

Chz1 is a histone chaperone that has been discovered to have the ability to bind H2A.Z[1]-H2B dimer and may play a role in the H2A.Z exchanging process. (Liu et al., 2014) Studies have found out about the structure and sequences of Chz1 (Figure 2). The Chz1 is usually an intrinsically disordered protein but can form a stable heterotrimer when interacting with H2A.Z-H2B. The binding protein of Chz1 has two α-helices on each end and a long irregular chain in the middle. The N-terminal of the binding protein (sequencing 81-93 of the whole Chz1 protein) binds to the α1 helix and L1 loop of H2A.Z. The long irregular chain in the middle (94-115) interacts with the α2 helices of H2A.Z and the α3 and αC helices of H2B. The C-terminal (116-132) binds to the α2 and α3 helices of H2A.Z (Zhou et al.,2008). The sequence of Chz1's binding protein has also shown the importance of recognizing the binding protein and the H2A.Z-H2B dimer, which is found to have rich D and E amino acids (Luk et al., 2007). Similar sequences have been found in other binding proteins.

YL1

The YL1 is a subunit of SRCAP[2], is critical in H2A.Z recognition, and can lead to transcriptional regulation (Xu M, 2021). The structure of YL1 has been found to rape the H2A.Z-H2B dimer and has three main binding sites. The first binding site is around the αC of H2A; the second is between the α1 and α2 helices of H2B, which is usually inferred as the H2B pocket; The third is around the H2A.Z's α1 and α2 (Figure 2). Moreover, the sequence of YL1 seems to follow the same pattern as Chz1(Liang et al. 2016)

ANP32e

ANP32e has been found to bind the H2A.Z-H2B dimer specifically and contributes to the deposition of H2A.Z (Mao et al., 2014). This process can regulate the transcription of DNA and influence DNA repair (Shin et al., 2018). The binding

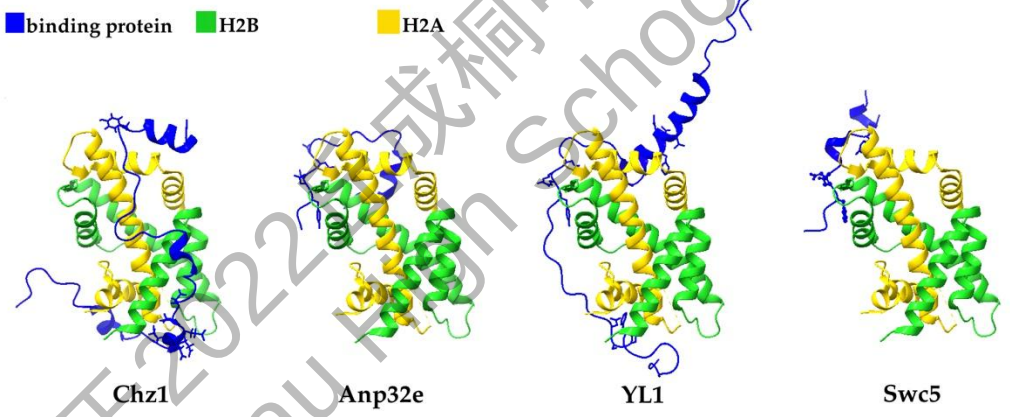
[1] histone H2A's variant, which is essential for specific genetic regulation and organism's development.
[2] SRCAP complex functioned through H2A.Z deposition to activate transcription of metabolic gene.

protein of ANP32e is at the C-terminal of ANP32e, and it contains a region rich in DE amino acids (Figure 2). The binding protein interacts with the L2 of H2A.Z and performs a firm binding (Obri et al., 2014).

Swc5

Swc5 is a subunit of the SWR complex[3] and has been found to bind to H2A-H2B with its intrinsically disordered region. The mutation in Swc5 in the specific binding sequence can inhibit the nucleosome editing function of the SWR complex. The short helix of the binding protein can interact with the L2 in H2A and L1 in H2B (Figure 2). The short helix has the same sequence that contains D, E, F, and Y as the other binding protein (Huang et al., 2019).

Other binding proteins such as Nap1 (Steven et al., 2003), FACT(Bhakat & Ray, 2022), and APLF (Corbeski et al., 2018) have also been found to have some similarities with the proteins stated above, most notably in the binding region and its binding sequence.



Chz1	TVEDSESDMDDAKLDALMGNEGEEEEDDLAEIDTSNIITSGRRTRGKVIDYKKTAEELDKKE
Anp32e	EVGLSYLMKEEIQDEEDDDDYVE
YL1	AASRSRRNNAGNKIAHLLNEEEEDDFYKTSYGGFQEDEEDKEYEQKDEEDVVDSDFSIDENDEPVSD
Swc5	DEDEDGYIEEEDEDFQPE

Figure 2 The binding proteins' structure and interaction with H2A-H2B dimer.

The yellow colored ones are the H2A; the green colored ones are the H2B; the blue colored ones are the binding proteins discussed. The atoms shown are the sequences that have been labeled blue in the chart below, which are the DEF/Y motif.

[3] SWR complex edits the histone composition of the nucleosome at promoters to facilitate transcription by replacing the two nucleosomal H2A-H2B dimers with H2A.Z-H2B dimers.

C. Characteristics of H2A-H2B binding protein

From the proteins shown above, we can find those binding proteins have specific characteristics in their structure and sequences.

The binding protein structure appears most likely intrinsically disordered (Warren & Shechter, 2017). The intrinsically disordered region (IDR) does not have a secondary structure and does not fold spontaneously into a single stable structure. Therefore, the flexibility and dynamics of the IDR are essential for histone binding.

Another structural characteristic of binding protein is that it contains three main binding sites, which are found according to the known histone binding proteins. In addition, the interaction at the three binding sites is most stable and plays a vital role in raping the H2A-H2B dimer (Figure 1B). The first binding site (Bd1) is at the αC and α3 helices of H2A. The second binding site (Bd2) is around the L1 of H2B and is the most common binding site found in most binding proteins. Finally, the third binding site (Bd3) is near the αN helix and α1 helix of H2A (Hammond et al., 2017). Although the Bd1 and Bd3 are not as widespread as Bd2, the two locations also contribute to the firm and stable engagement between binding protein and H2A-H2B dimer.

The sequences of amino acids may also play a part in the binding process. The sequence at the three binding sites has got a DEF/Y motif, which suggests a D/E-rich region, and starting with a D/E amino acid (this D/E amino acid is at position 1), there will be another D/E amino acid at position 2 or 3, followed by an F or Y amino acid. In other words, the sequence should show a DXXEF/Y or a DXEF/Y pattern, where D and E can be converted to each other (Huang et al., 2020). Moreover, the acidic nature, thus the continuous DE sequence, is used to bind H2A-H2B. The acidic amino acids tend to bind more tightly with the basic amino acid on the surface of H2A-H2B histone (Laskey et al., 1978). Through more research, another type of binding motif has also been found, which contains a W amino acid in the place of what was before known as F or Y. Further studies have demonstrated that the W amino acid is usually followed by a P amino acid (Mehrotra et al., 2010).

As we can see above, the structure and function of H2A-H2B binding proteins vary, and it is hard to find all of the binding proteins in a human's genome. Because of this, we wish to find a way to screen and identify the binding protein of H2A-H2B on a large scale and find the proteins that have been neglected by summarizing the features of binding proteins. Therefore, in this research, we have screened for the DNA

replication performing protein to test our method of finding the H2A-H2B binding protein.

2. Material and Method

A. Computer screening

The first part of the protein screening process is done by computer analysis, narrowing the protein number down through the proteins' function, sequence, structure, and predicted interaction with the H2A-H2B dimer. The protein sequences that appear to have the most substantial possibility will later be tested through biological experiments.

Firstly, a program was used to search for DNA replication proteins in the UniProt database[4]. After that, each protein sequence is scanned to see if it contains the DEF/Y motif. Moreover, although some proteins are related to humans, they are only from viruses or bacteria that can infect humans, so we eliminated the non-human proteins. Because the protein number after the three steps above is still too significant for analysis, a triple D or E motif is added to filter out the proteins that are not acidic and is hard to bind the histone. Finally, the qualified proteins are downloaded for the subsequent processes.

The UniProt database is then used to see the structure near the DEF/Y motif, and the sequences that are in the IDR are recorded in about 30 to 55 amino acids. There are three ways to record sequences: placing the DEF/Y motif in the N-terminal, the C-terminal, or the middle. The amino acids determine the way to record near the DEF/Y motif. In addition, the goal is to have more D/E acidic amino acids in the recorded sequence. If the acidic amino acid is evenly spread, we tend to place the DEF/Y motif near the N-terminal to have a higher possibility of getting a positive result in the latter interaction prediction screening. This is because the binding protein known all tend to rape the H2A-H2B dimer counter-clockwise, meaning the N-terminal of the binding protein is near the Bd2, the most common and stable binding site.

[4] https://www.uniprot.org/

The last part of computer screening uses alpha-fold (Jumper et al., 2021)(Evans et al., 2021) to predict the interaction between H2A-H2B dimers, and the result is then analyzed using ChimeraX (Goddard et al., 2017)(Pettersen et al., 2020). The two main characteristics of the structural binding, its direction and the binding sites, are evaluated. First, the direction of the binding protein should be counter-clockwise, meaning that the N-terminal is around the Bd1 and Bd2 while the C-terminal is around the Bd3. The second characteristic is if the DEF/Y motif is interacting near the three binding sites. During the analysis, the Bd2 is considered the most potent binding site, and the ones that do not bind to the Bd2 with its DEF/Y motif are eliminated.

After the processes above, the protein sequences that are most likely to be binding proteins and interact with H2A-H2B dimer are chosen for the following experiment.

B. Protein expression and purification

Transformation. We modified the existing pET-28a-based plasmid, which encodes an N-terminal hexahistidine (His6) tag and thrombin protease cleavage site followed by a SUMO tag. Because the plasmids of the proteins chosen above are lyophilized, they are first centrifuged at 1200 RPM for 1 min, and 20μL TB buffer is added to dissolve the plasmids. The plasmids for H2A-H2B are already in the form of a solution and can be used without the procedure above. 5μL of each plasmid solution is added to the thawed E. coli BL21 (DE3) cells and incubated in the mixture on ice for 20min. The tube is then heat shocked in a 42 °C water bath for the 60s to 90s and incubated on ice for another 2min. 500μL nonantigenic LB culture medium are added to the plasmids and shaken at 37 °C 250 RPM for 30min to 60min. After incubation, 100μL of each mixture is spread on the selection plate with kanamycin and grown overnight.

Protein expression. The bacteria grown in the selection plate are moved to an 800mL liquid LB medium with kanamycin and chloromycetin and shaken at 25°C for 90min for the OD_{600} to reach 0.6 to 0.8. 0.5mM of IPTG is then added to the medium, and the temperature is lowered to 15°C. The proteins are expressed overnight. The bacteria mixture is centrifuged at 4°C, 4000 RMP for 30min, and the sediment is collected.

Ni-NTA Purification. After the bacteria is harvested, it is resuspended using buffer A (10mM Tris 8.0, 2M NaCl, 20mM imidazole). The bacteria are then lysed through high pressure, and the mixture is centrifuged in a cryogenic superspeed centrifuge at 15000RMP for 90min, and the supernatant is collected. The nickel beads are first equilibrated using buffer A and then added to the supernatant to incubate at 4°C,

mixing continuously for 90min. The mixture is then poured into a gravity-flow column and use buffer B (20mM Tris 8.0, 2M NaCl, 30mM imidazole) to wash off the miscellaneous protein. CBB (Coomassie brilliant blue) checks if the miscellaneous proteins have been washed completely. After that, buffer C (20mM Tris 8.0, 400mM NaCl, 250mM imidazole) is used for the elute. Again, CBB is used to check if all the target proteins have been collected.

Ion exchange chromatography for H2A-H2B. In this study, an automated HPLC (high performance liquid chromatography) system was used for purification. First, the nickel-purified protein was desalted and deimidazoled. Then the sample was loaded into the substantial cation exchange column and balanced with low salt according to the negative charge of the protein. The high salt was combined to gradually change the ion binding strength through the gradient salt mixing procedure and separate the target protein from the miscellaneous proteins. Finally, the proteins in the collection tubes were selected according to the elution peak.

C. Protein affinity testing

His-tag digestion. The proteinase thrombin is used to recognize and cut off his-tag through the thrombin site. The plasmid used for the protein expression contains a thrombin site between the his-tag and SUMO tag, which allows us to cut off the his-tag from the binding protein for the pulldown experiment. 4μL of thrombin is added to the purified binding protein, and the mixture is incubated overnight at 4°C.

His-tag elimination through Ni-NTA. The mixture is then incubated with the nickel beads as the procedure stated above and then poured into a gravity-flow column. Because the binding proteins that do not have his-tag will not be able to bind with nickel beads, while thrombin and the his-tags that have been cut off will stay on the beads, the flowthrough and wash are collected.

Pull-down experiment. The H2A-H2B is first incubated with nickel beads in buffer A for 120min at 4°C. 50μL of the his-tag-H2A-H2B beads are then incubated with the binding protein (secondary protein) for another 120min. The mixture is washed with buffer B (20mM Tris 8.0, 2M NaCl, 30mM imidazole) to test the binding affinity of the two proteins. After that, buffer D (20mM Tris 8.0, 300mM NaCl, 250mM imidazole) is used for the elute. The mixture is then centrifuged, and the supernatant is run through the SDS-PAGE. Proteins were visualized by staining with CBB.

3. Result

A. The computer predictions of DNA replication-associated protein

These are the number of proteins remaining after each stage of the screening process: there are more than 6000 proteins that are associated with DNA replication, and more than 3000 of them contain the DEF/Y motif. After eliminating the non-human protein, the number decreases to about 2000+, and only about 300 of them show a strong acidity. 118 protein sequence is in the IDR. After analyzing the interaction between the predicted protein and H2A-H2B dimer, four binding proteins were selected for further experiments.

CHD1

The first protein is CHD1, which has been identified as an ATP-dependent chromatin-remodeling factor that functions as a substrate recognition component of the transcription regulatory histone acetylation complex (Flanagan et al., 2005)(Sims et al., 2007). It has also been found to regulate DNA replication negatively. Moreover, it is required to maintain a specific chromatin configuration across the genome and associate histone deacetylase activity. There have already been studies suggesting that CHD1 can bind with histone H3. Human CHD1 double chromodomains are involved in targeting the lysine 4-methylated histone H3 tail recognition (Sims et al., 2005). Although the sequence found in the previous study is not the same as the sequence corresponding to the interaction with H3, the predicted structure still shows a high possibility that CHD1 can bind H2A-H2B dimer. The binding protein of CHD1 interacts with the Bd2 and Bd3 of the H2A-H2B dimer through its “DEDW” motif and its “ESDY” motif, respectively (Figure 3).

NP1L1

The second binding protein is NP1L1, which has been studied to be a histone chaperone that participates in the nuclear import of H2A-H2B and nucleosome assembly (referred to as Nap1 in the passage above). In addition, it engages in the control of DNA-templated processes such as replication and DNA damage repair (Venkatesh & Workman, 2015). It has already been found that the C-terminal acidic domain of Nap1 contributes to its stability of histone binding (Andrews et al., 2008). We have predicted the interaction between NP1L1 and H2A-H2B dimer from the computer analysis and screening. The “DDDY” motif and the “DPDY” motif bind with the Bd2 and Bd3. The binding protein sequence of Nap1 has already been found

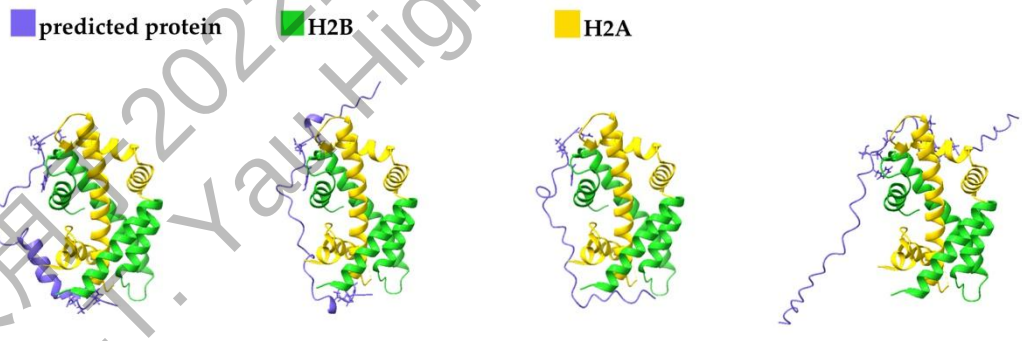
before, and it is highly conserved (only a few amino acids longer) to the sequence found in our computer screening (Ohtomo et al., 2016).

TCAB1

TCAB1 is an RNA chaperone that maintains the telomere structure (Chen et al., 2018). Other studies suggest that it is an essential component of the telomerase holoenzyme complex, which participates in the replication of chromosome termini that elongates telomeres in most eukaryotes (Venteicher et al., 2009). In addition, it can recruit the ubiquitin ligase RNF8 to break the DNA and promote both homologous recombination and non-homologous end joining (Henriksson et al., 2014). The predicted interaction between TCAB1 and the H2A-H2B dimer shows a nicely raped shape with Bd2, which contains an "ETEF" sequence (Figure 3).

H3BQ83

This protein has not been studied much, but it has been indicated to be essential in controlling DNA replication and maintaining replication fork stability (Zody et al., 2006). The predicted structure of H3BQ83 has shown a Bd1 and Bd2 in the interaction with H2A-H2B, which is the only protein found in the screening process to have Bd1, giving it a possibility to be H2A.Z-H2B binding protein (Figure 3).



CHD1	DEDWQMSGSGSPSQSGSDSESEEEREKSSCDETESDYE
NP1L1	YFTGEAIEDDDDDYDEEGEEADEEGEEEGDEENDPDYD
TCAB1	EGDPVSTPLETEFGSPSELSPRIEEQELSENTSLPAEE
H3BQ83	MLEPQENGVIDLPDYEHVEDETFPPEPPPASPERQDGEGTEPDEE

Figure 3 The structural interaction between predicted proteins and H2A-H2B.

Colored as illustrated, and the atoms shown are the sequences that have been labeled blue in the chart below, which are the DEF/Y motif.

B. SDS-PAGE and data analysis for purified protein

After being purified, the predicted proteins and H2A-H2B dimers went through SDS-PAGE and Coomassie blue staining. All of the five proteins have been shown to express, and they appear to be of a high degree of purity.

The proteins had a high expression in the condition provided, according to the first protein that had been purified, NP1L1. Although some of the target protein is lost in flowthrough and wash, there are still enough proteins remaining on the beads to be collected for the later experiment. The result for NP1L1 protein has two bonds, but because beta-mercaptoethanol is not added when running the SDS-PAGE, the disulfide bond may change the protein's size. Nevertheless, indicated by the size of the two bands, which shows a twofold relationship, the purification is still convincing. Moreover, the pure H2A-H2B proteins, meaning that there is only one band from the gel, are collected for the later experiment. Most of the bands after the ion exchange chromatography have shown a high purity and high yield. Finally, the remaining three proteins, CHD1, TCAB1, and H3BQ83, have also shown to be expressed and purified.

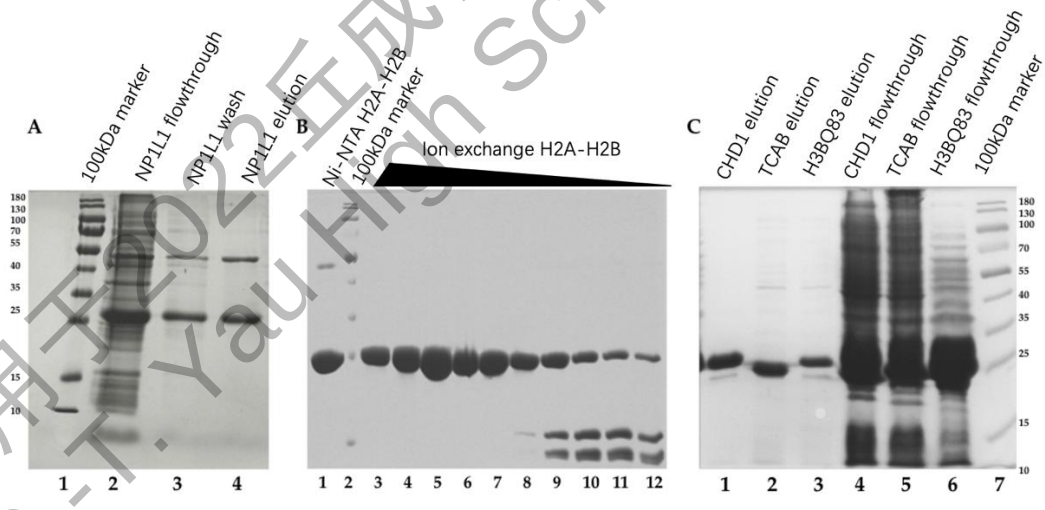


Figure 4 SDS-PAGE gel for binding protein and H2A-H2B.

(A) Coomassie-stained gel for NP1L1 purification. **(B)** The purification of H2A-H2B, both Ni-NTA purification and ion exchange chromatography. The 3 to 12 lines indicate the collections of protein during ion exchange chromatography, and the 3 to 8 lines (having a high purity) are stored for the latter experiment. **(C)** Coomassie-stained gel for the purification of CHD1, TCAB, and H3BQ83.

C. Predicted protein and H2A-H2B binding affinity

Because of the time limit, NP1L1 and TCAB1 are not tested for binding affinity, but NP1L1 has already been studied for binding with H2A-H2B.

NP1L1 has not been tested for affinity in this research because of the failed result during the incubation of the pull-down experiment (Figure 5). Fortunately, a previous study has already proven the binding of the particular sequence of NP1L1. The NP1L1 binding protein has been shown to have a strong endothermic interaction with H2A-H2B protein, and a mutant in this sequence will dramatically reduce its binding affinity with H2A-H2B. The CTAD is referred to as a binding protein in this passage (Hideaki et al., 2016).

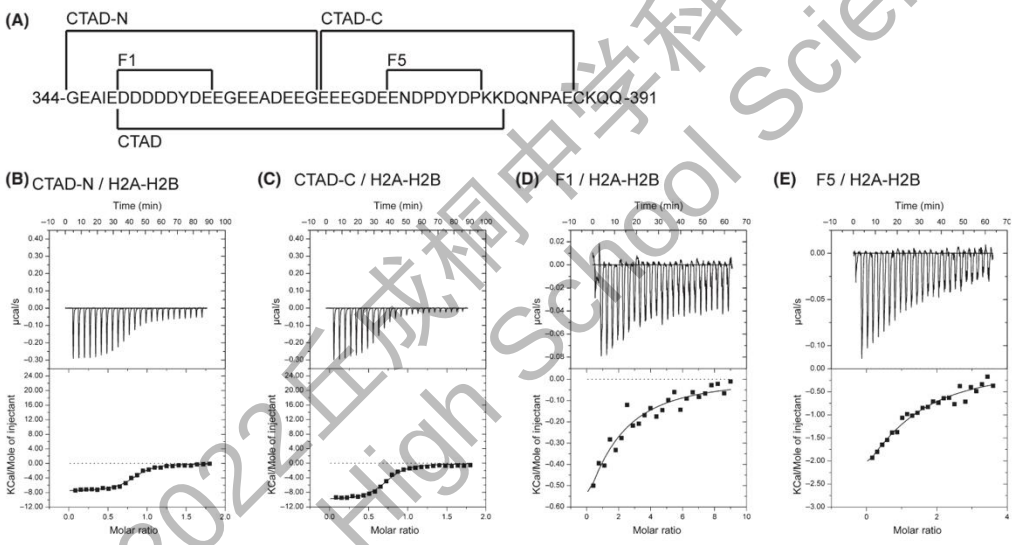


Figure 5 ITC binding curves for the interaction of NP1L1 and H2A-H2B.

(A) The amino acid sequence of NP1L1. **(B-E)** The interaction of NP1L1 with H2A-H2B. The top panels show the raw data, and the bottom panels show the fitted binding isotherms.

CHD1 and H3BQ83 have been tested through a pull-down experiment, and the Coomassie-stained gel has shown a positive result for both proteins. The H2A-H2B dimers are attached to nickel beads by his-tag, and the binding proteins are attached through the firmly binding between them and H2A-H2B dimers. Two bands of proteins have been reviled from the gel, which is the substance that is being eluted (Figure 6). This can prove that CHD1 and H3BQ83 do have a binding interaction.

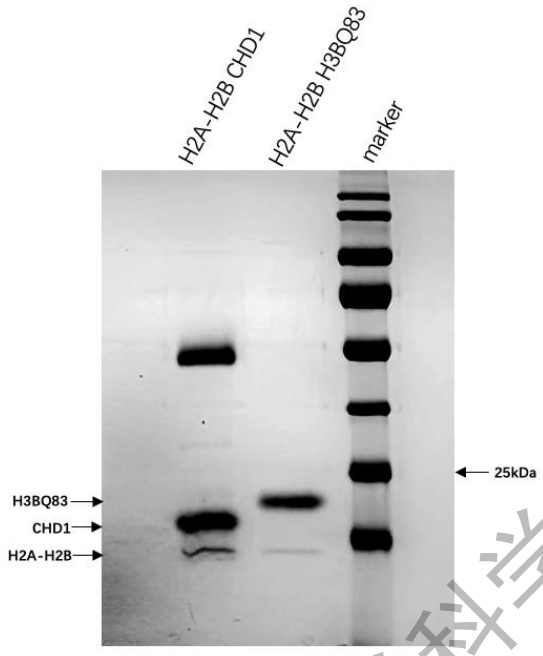


Figure 6 Coomassie-stained gel for pull-down experiment.

The purified binding proteins CHD1, H3BQ83, and H2A-H2B dimer were incubated with his-tag or his-tag-CHD1 or his-tag-H3BQ83 fusion protein bond to nickel beads. Bound proteins were separated by 12% SDS-PAGE and stained by Coomassie. Bindings of CHD1, H3BQ83, and H2A-H2B dimers were shown, respectively.

4. Discussion

A. Comparison between computer analysis and experiment data

The result of the computer analysis and biological experiment have been summarized (Table 1). Except for TCAB, all of the proteins predicted by computer analysis to bind with H2A-H2B have also shown binding affinity in the biological experiment.

The proteins had confirmed different characteristics of binding protein characteristics that are summarized in this research. In this study, the IDR and DEF/Y motif has been considered to be essential elements for H2A-H2B binding. The triple DE motif has effectively narrowed the range of proteins from the acidity characteristic. All the predicted binding proteins matched these characteristics, and previous studies have shown their importance. The correspondence of computer analysis and experiment confirmed the incommensurability of deletion of these screening conditions.

The binding features are the most visualized screening process and have shown the binding affinity of each predicted binding protein by the structural calculation of alpha-fold technology. Because of the limited time, quantitative research like ITC (Isothermal Titration Calorimetry) has not been done, and the binding intensity cannot be tested. However, alpha-fold analysis has dramatically helped to eliminate the low probability results as well as to help us come up with a more clear binding strategy, the structural interaction between binding protein and H2A-H2B dimer, and have a better understanding of the modification of H2A-H2B. The Bd2, previously detected as the most common, can be predicted by alpha-fold technology. The other two binding sites, the Bd1 and Bd3, may have also increased the affinity of H2A-H2B attachment.

In future research, quantitative research can supplement the alpha-fold technology to see how much more significant interaction is provided when Bd1 or Bd3 is present. Again, the large band at the top of the CHD1 gel is a result of disulfide bond.

B. The sufficiency of the screening method

The human proteome is massively huge, and the ones related to histone modification are enormous. This screening method used the proteins that function in DNA replication as a sample containing about 5000 proteins. These 5000 proteins are screened and narrowed down into four protein sequences that have been predicted by this method to have the ability to bind H2A-H2B dimer. Three of the four proteins have been successfully tested, and all three have shown to bind with H2A-H2B, proving the sufficiency of the screening method.

The core of the method is that a standard characteristic of binding protein has been reasonably summarized. Reducing the work to find new binding proteins. For example, the CHD1 has previously been shown to bind with histone H3 and has also been found, in this case, to bind H2A-H2B. Moreover, other proteins that have been neglected can also be found using this method. H3BQ83 is an example of a protein that has not been put into attention, but from the pull-down experiment, it has been shown to bind with H2A-H2B, and its Bd1 is also a strong suggestion that it may be an H2A.Z chaperone responsible for the exchange of H2A and H2A.Z.

Summary of computer analysis and biological experiment results

process		Np1L1	CHD1	TCAB1	H3BQ83
Computer screening features	DEF/Y motif	+	+	+	+
	IDR	+	+	+	+
	Triple DE motif	+	+	+	+
	Bd1	-	-	-	+
	Bd2	+	+	+	+
	Bd3	+	+	-	-
Biological experiment		+	+	\	+

Table 1 the "+" indicates a positive result, "-" indicates a negative result, and "\" indicates the data is missing.

C. Method optimization

To increase simplicity and sufficiency, we have designed certain optimizations, which increase the silico screening process and give more standardized criteria in the future.

First of all, the identification of protein IDR in this study was purely manual, and the structure of the sequence position given in the database was screened by visual observation. This process takes a long time and is less accurate than computer calculation. Therefore, an alternative to the manual screening method is very important. SlimSearch database provides conditions for the optimization of IDR identification (Krystkowiak & Davey, 2017). The website has a section called "Disordered Score Cut Off." If the score is greater than 0.4, the protein is considered to be IDR. In this way, the screening efficiency can be improved.

After doing quantitative studies and knowing the strength of each binding, a parameter can be set for each characteristic. For example, Bd2, being very important, can have a score of 50, while Bd1 and Bd3 are less common and will have a score of 30. The parameters can make the screening process more standardized and reduce bias.

Moreover, qualitative analysis of the binding strength of binding protein can further help us clarify the importance of each feature of binding protein in the screening process. The screening features satisfied by binding proteins with higher binding strength can be considered more important, while some features satisfied by predicted proteins that are actually found to be false positives after identification may be considered not important. Therefore, we can assign points to different features according to their importance. For example, Bd2, which has been identified as the most common binding site in previous studies and has shown positive results in numerous subsequent screenings, could be assigned a point of 50. However, Bd1 may also be positive if it is not satisfied, but meeting Bd1 can enhance the binding strength of the binding protein. Therefore, Bd1 can be assigned 20 points. Finally, the predicted protein with the highest point will be considered to have the highest possibility of being binding protein. The points can make the screening process more standardized and reduce bias.

5. Conclusion

This research developed an approach to screening and identification of H2A-H2B binding proteins. The previous researches about this type of protein vary in structure and function, making the identification of binding protein verbose. Six features of the binding proteins have been summarized in this study to develop a simple and effective method to fix this problem.

The H2A-H2B binding proteins have been found to have a DEF/Y motif and are present in IDR with high acidity. Their interactions with the H2A-H2B histone dimer have been found to have three main binding sites. All the characteristics stated above are included in this study to screen the binding proteins. The pull-down experiment is used to verify the computer screening process and identify the binding proteins. The data from the silico analysis and the experiment have been found to be highly conserved in this research. NP1L1, CHD1, and H3BQ83 are identified as binding proteins.

More studies can be further conducted to improve the present screening process, such as automatic IDR identification and quantitative experiment for binding affinity.

Reference

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, *389*(6648), 251–260. https://doi.org/10.1038/38444

Huang, Y., Dai, Y., & Zhou, Z. (2020). Mechanistic and structural insights into histone H2A–H2B chaperone in chromatin regulation. *Biochemical Journal*, *477*(17), 3367–3386. https://doi.org/10.1042/bcj20190852

Liu, H., Luo, K., Zhou, Z., Mu, Y., & Wan, Y. (2014). Histone chaperone CHZ1 facilitates the disfavouring property of SPT16 to H2A.Z-containing genes in *saccharomyces cerevisiae*. *Biochemical Journal*, *460*(3), 387–397. https://doi.org/10.1042/bj20140186

Dronamraju, R., Ramachandran, S., Jha, D. K., Adams, A. T., DiFiore, J. V., Parra, M. A., Dokholyan, N. V., & Strahl, B. D. (2017). Redundant functions for NAP1 and CHZ1 in H2A.Z deposition. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11003-8

Luk, E., Vu, N.-D., Patteson, K., Mizuguchi, G., Wu, W.-H., Ranjan, A., Backus, J., Sen, S., Lewis, M., Bai, Y., & Wu, C. (2007). CHZ1, a nuclear chaperone for histone H2AZ. *Molecular Cell*, *25*(3), 357–368. https://doi.org/10.1016/j.molcel.2006.12.015

Xu, M., Yao, J., Shi, Y., Yi, H., Zhao, W., Lin, X., & Yang, Z. (2021). The SRCAP chromatin remodeling complex promotes oxidative metabolism during prenatal heart development. *Development*, *148*(8). https://doi.org/10.1242/dev.199026

Liang, X., Shan, S., Pan, L., Zhao, J., Ranjan, A., Wang, F., Zhang, Z., Huang, Y., Feng, H., Wei, D., Huang, L., Liu, X., Zhong, Q., Lou, J., Li, G., Wu, C., & Zhou, Z. (2016). Structural basis of H2A.Z recognition by SRCAP chromatin-remodeling subunit YL1. *Nature Structural & Molecular Biology*, *23*(4), 317–323. https://doi.org/10.1038/nsmb.3190

Mao, Z., Pan, L., Wang, W., Sun, J., Shan, S., Dong, Q., Liang, X., Dai, L., Ding, X., Chen, S., Zhang, Z., Zhu, B., & Zhou, Z. (2014). ANP32E, a higher eukaryotic histone chaperone directs preferential recognition for H2A.Z. *Cell Research*, *24*(4), 389–399. https://doi.org/10.1038/cr.2014.30

Shin, H., He, M., Yang, Z., Jeon, Y. H., Pfleger, J., Sayed, D., & Abdellatif, M. (2018). Transcriptional regulation mediated by H2A.Z via ANP32E-dependent inhibition of protein phosphatase 2A. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1861*(5), 481–496. https://doi.org/10.1016/j.bbagrm.2018.03.002

Obri, A., Ouararhni, K., Papin, C., Diebold, M.-L., Padmanabhan, K., Marek, M., Stoll, I., Roy, L., Reilly, P. T., Mak, T. W., Dimitrov, S., Romier, C., & Hamiche, A. (2014). ANP32E is a histone chaperone that removes H2A.Z from chromatin. *Nature*, *505*(7485), 648–653. https://doi.org/10.1038/nature12922

Huang, Y., Sun, L., Pierrakeas, L., Dai, L., Pan, L., Luk, E., & Zhou, Z. (2020). Role of a DEF/Y motif in histone H2A-H2B recognition and nucleosome editing. *Proceedings of the National Academy of Sciences*, *117*(7), 3543–3550. https://doi.org/10.1073/pnas.1914313117

McBryant, S. J., Park, Y.-J., Abernathy, S. M., Laybourn, P. J., Nyborg, J. K., & Luger, K. (2003). Preferential binding of the histone (H3-H4)2 tetramer by NAP1 is mediated by the amino-terminal histone tails. *Journal of Biological Chemistry*, *278*(45), 44574–44583. https://doi.org/10.1074/jbc.m305636200

Bhakat, K. K., & Ray, S. (2022). The facilitates chromatin transcription (FACT) complex: Its roles in DNA repair and implications for cancer therapy. *DNA Repair*, *109*, 103246. https://doi.org/10.1016/j.dnarep.2021.103246

Corbeski, I., Dolinar, K., Wienk, H., Boelens, R., & van Ingen, H. (2018). DNA repair factor APLF acts as a H2A-H2B histone chaperone through binding its DNA interaction surface. *Nucleic Acids Research*, *46*(14), 7138–7152. https://doi.org/10.1093/nar/gky507

Warren, C., & Shechter, D. (2017). Fly fishing for histones: Catch and release by histone chaperone intrinsically disordered regions and acidic stretches. *Journal of Molecular Biology*, *429*(16), 2401–2426. https://doi.org/10.1016/j.jmb.2017.06.005

Hammond, C. M., Strømme, C. B., Huang, H., Patel, D. J., & Groth, A. (2017). Histone chaperone networks shaping chromatin function. *Nature Reviews Molecular Cell Biology*, *18*(3), 141–158. https://doi.org/10.1038/nrm.2016.159

Laskey, R. A., Honda, B. M., Mills, A. D., & Finch, J. T. (1978). Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature*, *275*(5679), 416–420. https://doi.org/10.1038/275416a0

Mehrotra, P. V., Ahel, D., Ryan, D. P., Weston, R., Wiechens, N., Kraehenbuehl, R., Owen-Hughes, T., & Ahel, I. (2011). DNA repair factor APLF is a histone chaperone. *Molecular Cell*, *41*(1), 46–55. https://doi.org/10.1016/j.molcel.2010.12.008

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., & Ferrin, T. E. (2017). UCSF Chimerax: Meeting modern challenges in visualization and analysis. Protein Science, 27(1), 14–25. https://doi.org/10.1002/pro.3235

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2020). UCSF chimerax: Structure visualization for researchers, educators, and developers. Protein Science, 30(1), 70–82. https://doi.org/10.1002/pro.3943

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2021). Protein complex prediction with Alphafold-Multimer. https://doi.org/10.1101/2021.10.04.463034

Flanagan, J. F., Mi, L.-Z., Chruszcz, M., Cymborowski, M., Clines, K. L., Kim, Y., Minor, W., Rastinejad, F., & Khorasanizadeh, S. (2005). Double Chromodomains cooperate to recognize the methylated histone H3 tail. Nature, 438(7071), 1181–1185. https://doi.org/10.1038/nature04290

Sims, R. J., Millhouse, S., Chen, C.-F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L., & Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors

and pre-mrna splicing. Molecular Cell, 28(4), 665–676. https://doi.org/10.1016/j.molcel.2007.11.010

Sims, R. J., Chen, C.-F., Santos-Rosa, H., Kouzarides, T., Patel, S. S., & Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. Journal of Biological Chemistry, 280(51), 41789–41792. https://doi.org/10.1074/jbc.c500395200

Venkatesh, S., & Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology*, *16*(3), 178–189. https://doi.org/10.1038/nrm3941

Andrews, A. J., Downing, G., Brown, K., Park, Y.-J., & Luger, K. (2008). A thermodynamic model for NAP1-histone interactions. Journal of Biological Chemistry, 283(47), 32412–32418. https://doi.org/10.1074/jbc.m805918200

Ohtomo, H., Akashi, S., Moriwaki, Y., Okuwaki, M., Osakabe, A., Nagata, K., Kurumizaka, H., & Nishimura, Y. (2016). C‐terminal acidic domain of histone chaperone humannap1 is an efficient binding assistant for histone H2A‐H2B, but not H3‐H4. Genes to Cells, 21(3), 252–263. https://doi.org/10.1111/gtc.12339

Chen, L., Roake, C. M., Freund, A., Batista, P. J., Tian, S., Yin, Y. A., Gajera, C. R., Lin, S., Lee, B., Pech, M. F., Venteicher, A. S., Das, R., Chang, H. Y., & Artandi, S. E. (2018). An activity switch in human telomerase based on RNA conformation and shaped by TCAB1. Cell, 174(1). https://doi.org/10.1016/j.cell.2018.04.039

Venteicher, A. S., Abreu, E. B., Meng, Z., McCann, K. E., Terns, R. M., Veenstra, T. D., Terns, M. P., & Artandi, S. E. (2009). A human telomerase holoenzyme protein required for cajal body localization and telomere synthesis. Science, 323(5914), 644–648. https://doi.org/10.1126/science.1165357

Henriksson, S., Rassoolzadeh, H., Hedström, E., Coucoravas, C., Julner, A., Goldstein, M., Imreh, G., Zhivotovsky, B., Kastan, M. B., Helleday, T., & Farnebo, M. (2014). The scaffold protein WRAP53β orchestrates the ubiquitin response critical for DNA double-strand break repair. Genes & Development, 28(24), 2726–2738. https://doi.org/10.1101/gad.246546.114

Zody, M. C., Garber, M., Sharpe, T., Young, S. K., Rowen, L., O'Neill, K., Whittaker, C. A., Kamal, M., Chang, J. L., Cuomo, C. A., Dewar, K., FitzGerald, M. G., Kodira, C. D., Madan, A., Qin, S., Yang, X., Abbasi, N., Abouelleil, A., Arachchi, H. M., … Nusbaum, C. (2006). Analysis of the DNA sequence and duplication history of human chromosome 15. Nature, 440(7084), 671–675. https://doi.org/10.1038/nature04601

Krystkowiak, I., & Davey, N. E. (2017). SLiMSearch: A framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Research, 45(W1). https://doi.org/10.1093/nar/gkx238

Acknowledgment

I am indebted to my two instructors, Dr. Zheng Zhou and Dr. Xiangmei Dou, and my two senior apprentices, Mr. Wencang Huang and Mr. Xin Yang. They taught me how to purify proteins and do pull-down experiments and helped me to design this research. Dr. Zheng Zhou helped me clarify the research direction and helped me decied the topic according to my ability. My instructors acted as gatekeepers through the whole research process and instructed me through my paper writing. Moreover, all of them taught me the spirit of doing scientific research. I also want to thank my parents and friend who supported me during my research and comforted me when the experiment failed.