Analysis of the Relationship Between Co-localization of Proteins in Protein Assemblies and Protein Proximity in 3-D Space

By:

Longsheng Xie

A Thesis
Submitted to the Faculty
Of

Worcester Polytechnic Institute

In partial fulfillment of the requirements for the

Degree of Master of Science

In

Bioinformatics and Computational Biology

September 2020

APPROVED:

_____
Dr. Dmitry Korkin, Major Advisor

_____
Dr. Elizabeth F. Ryder, Reader

Abstract:

There has been considerable interest in defining the complete set of human protein complexes. Currently, a comprehensive dataset of human protein complexes, hu.MAP (http://proteincomplexes.org) gives us a valuable source for a better understanding of the core cellular functions of human proteins and helping to know which proteins will work together(Drew et al., 2017). Recently, techniques allowing prediction of three-dimensional (3-D) chromosome and genome structures have become more mature. In particular, Hi-C data, which stores the interaction strength of different sites of each human chromosome, has become increasingly available. Therefore, based on hu.MAP and Hi-C data, we have information both about which genes are found in protein complexes, and the strength of interaction of each pair of genes in those human protein complexes at the level of their encoding DNA. In this work, we hypothesize that no matter whether each pair of interacting genes is located on the same chromosome or not, and no matter how far apart they are genome sequence-wise, when they encode proteins that ends up in the same protein complex, they will be physically closely located in 3-D space. Most of the genes that encode proteins that from the same complex were found to be not on the same chromosome, so we divided them into two groups: occurring either on the same chromosome or occurring on different chromosomes. For each pair of genes in a human protein complex, we randomly generated 100 pairs as a control group, then observed the distributions of the interaction strength of the experimental data and control group for the same chromosome, and the experimental data and control group for the different chromosomes. After that, for the gene pairs located on the same chromosome, we compared those gene pairs with the topologically associating domain (TAD). We found that TAD is the most critical factor affecting

the interaction between genes. Whether the interaction between proteins has an effect on the folding of chromosomes in the 3-D space requires a further in-depth study.

Acknowledgements:

Firstly, I would like to thank my thesis advisor Dr. Dmitry Korkin. The door to Prof. Dmitry office was always open whenever I ran into a trouble spot or encountered a question about my research or writing. He consistently encouraged me to complete this paper by myself individually but steered me in the right direction whenever he thought I needed it. Next, I am also thankful to Dr. Elizabeth F. Ryder for guiding me in every stage of this research paper. Without her support it would have been very difficult for me to prepare the paper to be meaningful and interesting. The two-year long journey of mine studying at WPI has made me fully understand and appreciate the field of Bioinformatics and Computational Biology. I want to thank all of the faculties I have been fortunate enough to learn from through my coursework. Finally, I would like to appreciate all my fellows BCB students and all Korkin lab-mates for their support and feedback throughout the past two-year time, simultaneously, this past two-year long studying and researching experience was really enjoyable and meaningful towards my entire life.

Table of Contents:

List of Figures

1. INTRODUCTION

Two or more associated polypeptide chains with different functions formed the protein complex. Within a protein complex, proteins link with non-covalent protein-protein interactions. Different protein complexes perform different functions, and the same complex can perform very different functions depending on many factors. The physical interaction of proteins is responsible for performing many of the essential functions of cells and organisms; mutations that destroy protein interactions can lead to loss of function and hence genetic diseases. Technical advances in the field of proteomics have witnessed the partial reconstruction of the protein interaction networks in humans and other animals. Moreover, with the first ongoing efforts made on the map of the human protein complex, we can better understand the composition, formation, and function of the majority of human protein complexes. (Drew, et al., 2017) Currently, there is a great interest in defining the relationship between the construction of the human protein complex and the chromosome in the 3-D space.

In eukaryotes, the genomes are not formed as a linear molecule entity but hierarchically located inside the nucleus which provide a confined space for folding DNA and enable genes to express at the right time and in the right place. The three-dimensional configuration of the chromosome modulates biological processes such as transcription, DNA replication, cell division, meiosis, protein expression (Oluwadare et al., 2019), which are crucial for cell differentiation and animal development. The three-dimensional folding of chromosomes compartmentalizes the genome and can bring distant functional elements, such as promoters and enhancers, into close spatial proximity(van Berkum et al., Cell, 2010). In this paper, we are trying to determine whether genes that express the same protein complex are located in the close physical proximity in their 3-D

chromosome folded structure, even though they may be located relatively far apart in the linear structure of a chromosome, or on different chromosomes. In addition, we are also interested in whether the presence of topologically associating domains(TADs) will impact the result (Figure 1.). As a self-interacting genomic region, DNA sequences within a TAD physically interact with each other more frequently than with sequences outside TAD. And TADs were discovered using chromosome conformation capture techniques including Hi-C.

Hi-C is a genome-wide sequencing technique that can be used in investigating 3D chromatin conformation inside the nucleus. Hi-C measures pairwise contacts between virtually any pair of genomic loci and finds the nucleotide sequence of interacting fragments using high- throughput sequencing. Based on the Hi-C data, there is a contact matrix with the value of interaction strength of each pair. The range of values is from -infinity to 0; the smaller the value, the weaker the interaction of a pair. Through paired end sequencing, the Hi-C technique can retrieve a short sequence from the end of each ligated fragment. Since the 3-D chromosome structure is fixed, and then fragments that are close together in the 3-D structure are ligated together. In this way, each obtained ligated fragment, the two sequences represent two different restriction fragments, which are ligated together in the proximity ligation step (Hakim O et al., Cell, 2012). The pair of sequences are individually aligned to the genome. Thus, the fragments involved can be determined in this process. Thus, the Hi-C technique allowed all possible pairwise interactions between fragments to be tested. And we used a result of a Hi-C experiment test, from 4DN data portal (https://data.4dnucleome.org/) which contains all values of the interaction strength of each fragment pair and the interaction strength correspond to how close together the gene pairs are in the 3-D structure. Computational methods analyze Hi-C data and identify genome interactions

and topologically associating domains from genome-wide contact probability maps. The purpose of our research was to use Hi-C data to explore whether genes that express the same protein complex have strong interaction, and location relationship in 3-D structure.

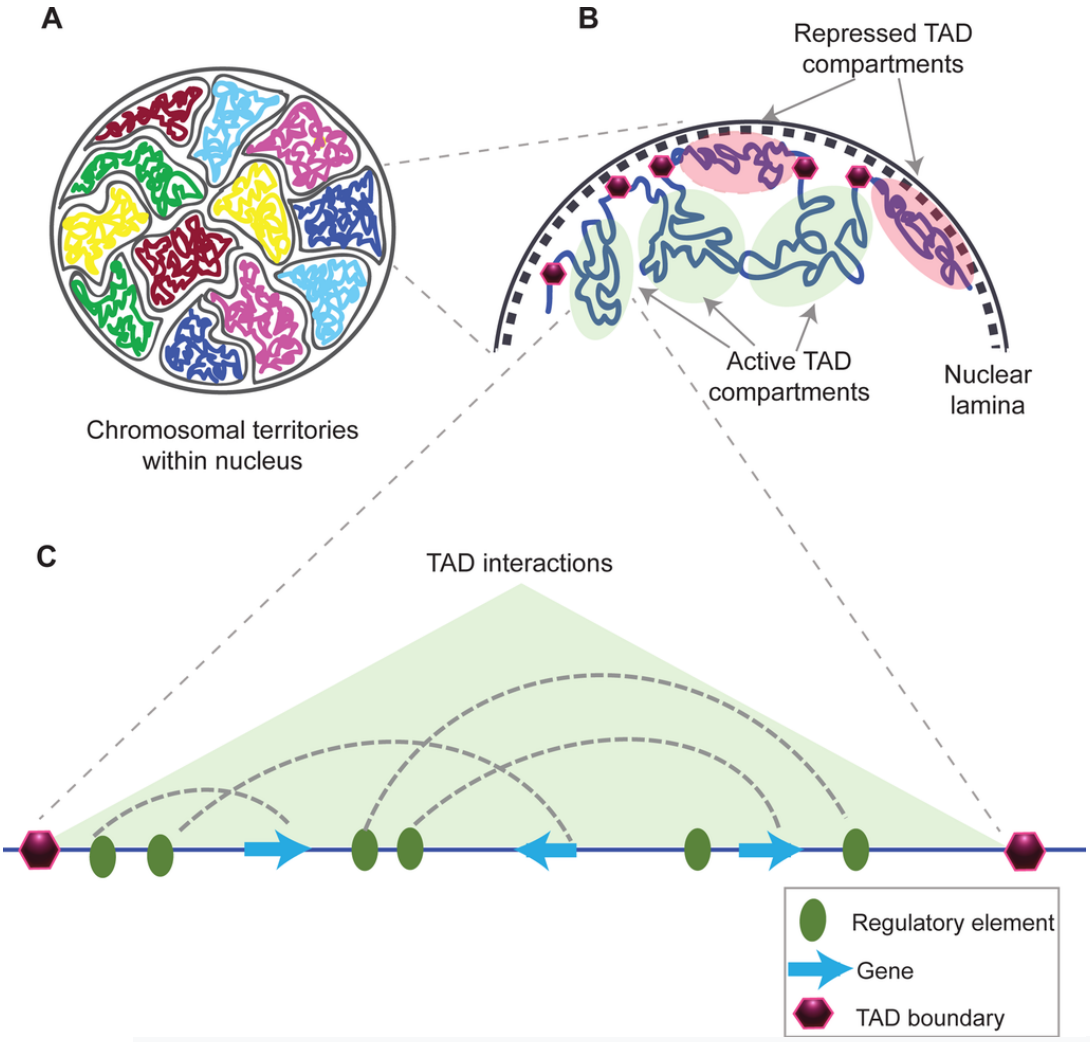

Figure 1. Topologically associating domains within chromosome territories, their borders and interactions (Source: https://en.wikipedia.org/wiki/Topologically_associating_domain)

2. METHODS

2.1 Problem Formulation

The goal of this project is to determine the extent to which the proteins that are functionally connected through participating in molecular assemblies are also geographically connected through their location proximity on a genome sequence and in 3-D space. Specifically, we hypothesize that when there is an interaction between proteins, the genes that encode these proteins, even if they are located on different chromosomes or at a greater distance on the same chromosome, these genes will also in close physical proximity due to the folding of chromosomes in the 3-D space.

To test this hypothesis, we selected groups of genes constituting a human protein complex and compared them with the genes located in random locations. We extracted the proximity of the genes on a chromosome by automatically retrieving their coordinates from Genome Ensembl and tested the strength of the interaction between genes using Hi-C data for both the genes located on the same chromosome or not. Finally, we compared genes encoding proteins that form protein complexes with the randomly selected gene pairs and for the gene pairs are located on the same chromosome, we also compared their interaction strength with the interaction strength of gene pairs within TADs.

2.2 Methodology Overview

To achieve our goal, we develop a fully automated computational pipeline(Fig. 1). The pipeline consists of five stages. First, we used human protein complex data from hu.MAP to do raw dataset processes. Next, we calculated distance between genes. At stage three, we calculated

interaction strength of each gene pair from human protein complex based on Hi-C data result. As the fourth step, we separated those gene pairs into two groups, one is on the same chromosome, the other is on the different chromosome, and made distributions for each group. Finally, we randomly generated control groups for each group and made the distributions.

Throughout the whole research process, except for the most commonly used Python libraries for data analysis (such as NumPy, Scipy, Matplotlib, and pandas), all coding work was performed in Python. In addition, there are some Python libraries for analyzing biological data (such as ensembl_rest and cooler) that were also be used in this research (Figure 2). The original data set of this research is derived from hu.MAP. Furthermore, Hi-C data is also important for us to get the strength of genes interaction. Besides these python libraries, the official website https://ensembl.org/ and https://data.4dnucleome.org/ also was used when the data is incorrect or incomplete in python libraries.

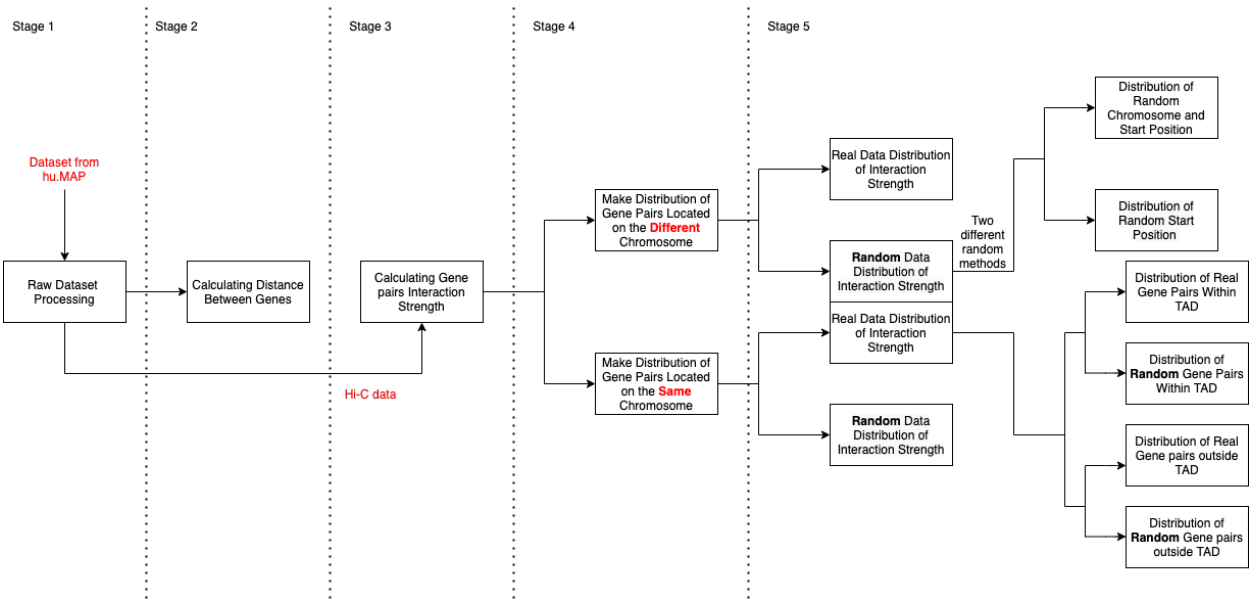

Figure 2. Pipeline of experiment methodology: The original data set of this research was obtained and extracted from hu.MAP, which contains 4,659 human protein complexes. We then reduced the dataset to 4,575 for further research, because we excluded some invalid data. Next, we calculated the distance between genes and interaction strength between gene pairs. Following this, we divided the data set into two groups, one with gene pairs on different chromosomes, and one with gene pairs on the same chromosome. For the gene pairs that are located on the same chromosome, we also compared their interaction strength with the interaction strength of gene pairs within TADs. As a control, we generated 100-fold more random data than our real data set.

2.3 Integration of Distance Between Genes

The raw dataset for carrying on this research is a set of the NCBI Entrez Genes identifiers and Ensembl identifiers, which is located in a human protein complex map containing 4,658 human protein complexes. The number of genes located in each human protein complex varies. All these genes encode proteins that interact with each other. Most of these gene identifiers are coded in NCBI Entrez Genes, and only tens of them are coded in the Ensembl. In order to unify all these gene identifiers, the Ensembl identifiers would be kept intact, with the implementation of the Python library MyGene, NCBI Entrez Genes identifiers could be retrieved in the library, and the first step of processing the raw data is to convert all these NCBI Entrez Genes identifiers

to Ensembl identifiers for further step research. Unfortunately, due to the rapid library update of the Ensembl identifiers and the python library may not catch up with the official Ensembl library update. Besides, with the Ensembl identifiers updated, the information could not be retrieved, some of the NCBI Entrez Genes identifiers is unable to correspond to one Ensembl identifier exactly, at the same time, some of the NCBI Entrez Genes would correspond to more than one Ensembl identifiers. In order to solve the above problems, all suspicious data are manually retrieved and converted according to the latest official library on the official website of ensembl.org. Formally, the raw data have been rearranged into a two-column data frame, the first column is the index number of the human protein complex in the raw dataset, this column could be treated as the index number and would be conveniently for a further step data analyzing, and the data in the second column is the Ensembl identifiers.

In this research stage, we applied a python library called ensembl_rest. By implementing this library, once the Ensembl identifier and other relevant attributes of the object are determined, the retrieved information can be derived. For our research purposes, we hope to gain the information about which chromosome the gene is located on, the start and end positions of the chromosome. Unfortunately, when searching, sometimes multiple Ensembl identifiers are retrieved, which means that the retrieved data contains contigs, that is, a group of overlapping DNA fragments together represent the common region of DNA. In addition, some Ensembl identifiers have been updated, and all original information has been moved to the new Ensembl identifiers. According to the statistical results of exception handling, less than 3% of the data have encountered this problem. We manually solved this problem by querying the Ensembl identifier on the official website of ensembl.org. After formatting, we finally obtained a 5-column data frame, which

contains the human protein complex index, the Ensembl identifier, the index number of the human chromosome, the starting position on the chromosome, and the ending position on the chromosome.

According to the combination theory, we use the letter n to represent the number of Ensembl identifiers in the human protein complex. When comparing any two Ensembl identifiers separately, a size table of n times n can be obtained. In this table, due to the symmetric relationship, we only fill in the general results. For one gene pair, we used the absolute difference of two genes' starting position as the distance difference. In one human protein complex, the number of gene pairs is a sum of an arithmetic progression. Finally, we can get an $n*(n-1)/2$ distance difference of all gene pairs distances in one human protein complex. Specifically, we want to compare the index numbers of the human chromosomes between a pair of genes. Once the index numbers of the human chromosomes are the same, the distance difference between the pair of genes is obtained by calculating the difference between the two genes. The difference between the two starting positions of a gene on a chromosome. In the subsequent research, for all calculations of distance difference, we have continued to use the method of starting position calculation. Also, when the genes are located on different chromosomes, the letter "I" will be marked in the table.

2.4 Integration of Interaction Strength Between Genes

In order to calculate the strength of interaction between genes on all human chromosomes, we used 4DN Data Portal that could provide an accompanying Jupyter Notebook Service for all end-users so that it is unnecessary for us to download extra-large files. We can select two ranges that

contains genes' starting position and ending positions. Then based on a Python library named cooler that was designed for the storage and manipulation of extremely large Hi-C datasets and 4DNFIB59T7NN.mcool file that contains a contact matrix of Hi-C experimental values, we will extract specific value that is the interaction strength between a gene pair from the contact matrix. Next, we calculated the log base 10 this specific value as the interaction strength we will use for the following step. Because after we take the logarithm, we might reduce skew to assist in modeling and straighten a nonlinear relationship in a scatterplot, so that we can model the relationship with simpler methods. We repeated this pipeline, then we can get all strength of interaction between a pair of genes. Based on the pairs of genes in the human protein complexes, we can, in turn, retrieve the interaction strength between each pair of genes. Formatting, like the table we obtained in the previous research phase, there exists several genes on a human protein complex, in this case, an n times n table was created after calculating the interaction strength between genes on a chromosome. However, unlike the previous research phase, the comparison of the index number on the human chromosome is unnecessary during this research phase. In other words, whether or not these pairs of genes are on the same human chromosome or not, we can retrieve the strength of their interaction. Then, based on the consideration of whether each pair of genes are located on the same chromosome or not, we decided to divide it into two categories for in-depth analysis.

Based on the dataset we obtained from the previous research phase, we separate the gene pairs tables into two groups, the group contains all the gene pairs which locates on the same human chromosome, in other words, they have the same index number of the human chromosome. And the second group contains all the gene pairs which are located on the different human

chromosomes, that is, they have different chromosome index numbers. Formally, these two data frames are four-column ones, the first and the third column represent the index number of the human chromosome, and the second and fourth columns represent the starting positions of the genes.

2.5 Comparison of Interaction Strength Between Random Gene Pairs and Complex Gene Pairs

2.5.1 Gene Pairs Located on Different Chromosomes

In this research stage, first, we collected all gene pairs on different chromosomes. Then, we based on a Python library named cooler and 4DNFIB59T7NN.mcool file to obtain the interaction strength between a pair of genes that are located on different chromosomes. After that, some results showed NaN and -Inf. In order to make the results more meaningful, we excluded all results showing NaN is shown which means there is no experiment test for this pair. After that, we calculated the reciprocal (because we kept the -Inf values which means really weak strength of interaction. -Inf values correspond to extremely weak interaction. In order to display and analyze a distribution of interaction values, we converted these values to reciprocals). We randomly generated the starting position of a pair of genes on different chromosomes 100 times and the index of the chromosome is the same with gene pairs in the real data set. Then, we calculated the interaction strength between the gene pairs for these random data. The data size was 5,511,300. After excluding all NaNs in the data set, we retained 4,214,000 and then calculated the reciprocal of all these data. Then we also analyzed the distribution for all data we gained in the next step. Finally, we randomly create the starting positions of gene pairs on any

two randomly created chromosomes. As in the previous step, another normal distribution analysis was performed.

2.5.2 Gene Pairs Located on Same Chromosomes

In this research phase, firstly, we recollected all the gene pairs on the same chromosomes, same to the previous research step, we have calculated the interaction strength between the gene pairs. Similar to the results in the previous research phase, we also obtained two distributions, one is based on the human protein complex, and the other is based on the random data. However, in this case, for each real gene pair located on a single chromosome, random gene pairs were located on that same chromosome, at the same distance apart as the real pairs.

As to the gene pairs located on the same chromosome in the human protein complex, we divided these gene pairs into two groups which locate in TADs(Topologically Associating Domain) or not separately, then exclude all the data shows NaN and get the reciprocal value, finally performed the normal distribution for these gene pairs in these two groups separately.

3. Results

Initially, we hypothesize that genes that encode the same protein complexes will be located in close physical proximity more frequently than the random data set we generated in 3-D structure. By using Hi-C data to collect strength of interaction of genes, the outputs indicated that the real data set was similar to the random data set, which means that genes encoding the same protein complex do not have a strong physical position relationship.

3.1 Raw Dataset Processing

The original dataset includes the genes that encode human protein complexes. Most of the gene IDs are encoded in the NCBI Entrez gene identifier, while a small number of gene IDs are encoded in the Ensembl identifier. Each row in the original data set represents genes contained in a human protein complex (Figure 3). In order to obtain the position of each gene on the chromosome, all gene IDs were converted to Ensembl identifiers. While using the Python library Mygene for conversion, what has been represented by the Ensembl identifier remains intact. Regarding some exceptions in the original data set, we manually retrieved these data on the official website ensembl.org. For some data that could not be retrieved, we excluded the entire human protein complex where the data is located (Figure 4.).

Complexes generated from two stage clustering of fully integrated protein interaction network

Complex Index	Gene ID	Gene ID	Gene ID	Gene ID	Gene ID	Gene ID	Gene ID
1	2712	23036	6504				
2	55623	800	91893				
3	138199	94059	4306	23098			
4	26190	139067	115290				
5	54414	149175					
6	342538	ENSG00000141048	10728				
7	55607	22853	54976	79834	81566	255057	9627
8	55165	51112					
9	163882	255057	81566	5501			
10	64651	5499	284352	339804			
11	22863	29982	387104				
12	2785	26747	23321	253559	5138	9326	192683

Figure 3. Raw dataset obtained from hu.MAP: This figure shows a partial raw dataset obtained from hu.MAP; the entire dataset size is 4,659 rows. Each row represents all of the genes encoding the proteins in a particular complex.

The result of converting gene id to ensemble id

complex	gene id	gene id	score	ensemble id		
(1	{'out': [{'query': '153129'	'_id': '153129'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000177058'}}]	'dup': []	'missing': []})
(1	{'out': [{'query': '10670'	'_id': '10670'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000155876'}}]	'dup': []	'missing': []})
(1	{'out': [{'query': '64121'	'_id': '64121'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000116954'}}]	'dup': []	'missing': []})
(2	{'out': [{'query': '441502'	'_id': '441502'	'_score': 0.5}]	'dup': []	'missing': []})	
(2	{'out': [{'query': '3024'	'_id': '3024'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000124610'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '2648'	'_id': '2648'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000108773'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '26009'	'_id': '26009'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000036549'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '10474'	'_id': '10474'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000171148'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '57325'	'_id': '57325'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000149474'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '8850'	'_id': '8850'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000114166'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '55689'	'_id': '55689'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000163872'}}]	'dup': []	'missing': []})
(3	{'out': [{'query': '6871'	'_id': '6871'	'_score': 1.55	'ensembl': [{'gene': 'ENSG00000276234'}	{'gene': 'ENSG00000277104'}]}]	'dup': [] 'missing': []
(4	{'out': [{'query': '3189'	'_id': '3189'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000096746'}}]	'dup': []	'missing': []})
(4	{'out': [{'query': '144983'	'_id': '144983'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000139675'}}]	'dup': []	'missing': []})
(5	{'out': [{'query': '22908'	'_id': '22908'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000211456'}}]	'dup': []	'missing': []})
(5	{'out': [{'query': '9276'	'_id': '9276'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000184432'}}]	'dup': []	'missing': []})
(5	{'out': [{'query': '26958'	'_id': '26958'	'_score': 1.55	'ensembl': {'gene': 'ENSG00000158623'}}]	'dup': []	'missing': []})

Figure 4. Partial table of Ensembl identifiers: The figure represents a partial result of processed data of Ensembl identifiers converted from the NCBI Entrez Genes identifiers.

3.2 Retrieving Ensembl Identifiers

By using the python library ensembl_rest, once the Ensembl identifier and other related attributes of the object were determined, we retrieved the relevant information of the gene, such as the first column with the same number belongs to one complex, the chromosome where the gene is located, and the start and end positions of the gene (Figure 5.). Regarding some exceptional genes, we cannot retrieve the Ensembl identifier, so we marked the “not found” symbol in its position; if the human protein complex contained at least one “not found” mark, this human protein complex was excluded from analysis.

The result of gene location

complex ensemble id chr position

80,('ENSG00000156017', 'chromosome:GRCh38:9:74980790:75028423:-1')
80,('ENSG00000275183', 'chromosome:GRCh38:19:54461732:54463778:-1')
80,('ENSG00000151623', 'chromosome:GRCh38:4:148078762:148444698:-1')
80,('ENSG00000004139', 'chromosome:GRCh38:17:28364356:28404049:1')
81,('ENSG00000119402', 'chromosome:GRCh38:9:120751978:120793416:-1')
81,('ENSG00000189252', 'chromosome:GRCh38:X:143508735:143517475:-1')
81,('ENSG00000269190', 'chromosome:GRCh38:19:38941401:38975742:-1')
82,('ENSG00000110013', 'chromosome:GRCh38:11:124633113:124695707:-1')
82,('ENSG00000185090', 'chromosome:GRCh38:1:37793802:37801137:1')
83,('ENSG00000253506', 'chromosome:GRCh38:17:61590421:61591219:-1')
83,not found
83,('ENSG00000110958', 'chromosome:GRCh38:12:56663341:56688408:-1')
84,('ENSG00000158528', 'chromosome:GRCh38:7:94907202:95296415:1')
84,('ENSG00000164715', 'chromosome:GRCh38:7:98106862:98209638:1')
84,('ENSG00000101220', 'chromosome:GRCh38:20:3753508:3768387:-1')
84,('ENSG00000173517', 'chromosome:GRCh38:15:77100656:77420144:-1')
84,('ENSG00000110925', 'chromosome:GRCh38:12:51061205:51083664:-1')
84,('ENSG00000099625', 'chromosome:GRCh38:19:1228287:1238027:-1')
84,('ENSG00000064692', 'chromosome:GRCh38:5:122311354:122464219:1')
84,('ENSG00000167615', 'chromosome:GRCh38:19:54448887:54462037:1')
84,('ENSG00000186298', 'chromosome:GRCh38:12:110719680:110742939:-1')

Figure 5. Partial table of Location of each gene: This figure shows a partial table of genes' location on a chromosome, with their beginning and ending base pairs. Human protein complexes were excluded from analysis manually if it contained at least one "not found" tag.

3.3 Calculating Distance Between Genes

When we use the alphabet n to indicate the number of Ensembl identifiers used to encode human protein complexes, we can plot an n times n table to show the difference in distance between any two genes that encode human protein complexes (Figure 6). First, we compared the index numbers of human chromosomes. When the index numbers of human chromosomes were the same (genes are found on the same chromosome), the distance difference between the two starting positions was calculated and recorded in the intersection table. When the index numbers of the human chromosome were different, we recorded an "I" mark in the intersection table. As a result, we calculated $n*(n-1)/2$ number of distance differences in a human protein complex,

where *n* is. In this table, because the two groups of genes are compared separately, there will be symmetry. The diagonal position, where genes are compared with themselves, are marked with "0", and the half of the table below the diagonal is left blank to avoid redundancy. Finally, we also conducted statistics to calculate how many "I" tags have been marked and how many "non-I" tags have been marked.

Complex0001	ENSG00000177058	ENSG00000155876	ENSG00000116954
ENSG00000177058	0	I	I
ENSG00000155876		0	I
ENSG00000116954			0

Complex0002	ENSG00000196933	ENSG00000124610
ENSG00000196933	0	I
ENSG00000124610		0

Complex0003	ENSG00000108773	ENSG00000036549	ENSG00000171148	ENSG00000149474	ENSG00000114166	ENSG00000163872	ENSG00000276234
ENSG00000108773	0	I	I	I	I	I	4706225
ENSG00000036549		0	I	I	I	I	I
ENSG00000171148			0	I	10260586	173917937	I
ENSG00000149474				0	I	I	I
ENSG00000114166					0	163657351	I
ENSG00000163872						0	I
ENSG00000276234							0

Complex0004	ENSG00000096746	ENSG00000139675
ENSG00000096746	0	I
ENSG00000139675		0

Complex0005	ENSG00000211456	ENSG00000184432	ENSG00000158623
ENSG00000211456	0	93666544	I
ENSG00000184432		0	I
ENSG00000158623			0

Figure 6. Partial table of distance differences of gene pairs in each complex: Each complex contains a number of genes ranging from 2 to 149. "I" indicates that the pair of genes are located on different chromosomes. The numbers indicate that the pair of genes are located on the same chromosome and represent the difference between the starting sites.

3.4 Calculating Interaction Strength

Similar to the previous research phase, we annotate n as the number of Ensembl identifiers for encoding a human protein complex, an n times n table could be drawn to show the interaction strength value obtained by using Jupyter Hub based on the implementation of the 4DN Data Portal between any two genes mutually for encoding a human protein complex (Figure 7.).

From the observation of Figure 6., we would like to have a deeper comprehension on the distribution situation of gene pairs' interaction strength, so we have created a histogram and compared it with a normal distribution (Figure 8a.). From this graph, we are able to observe that more than 75% of the data falls at -4 or below; more negative numbers indicate weaker interaction strength. Because the distribution of interaction strengths seemed to cluster into a group with very strong interactions and a group with very weak interactions, we hypothesized that these groups might correspond to genes located on the same and different chromosomes, respectively, with genes located on the same chromosome having a closer physical distance in 3-D structure. To test this hypothesis, we divided the gene pairs into two groups by a filtering condition that determined whether they are located on the same chromosome.

Complex0001	ENSG00000177058	ENSG00000155876	ENSG00000116954
ENSG00000177058	-0.745387602	-4.626067204	-4.431186732
ENSG00000155876		-0.830909415	-4.825142443
ENSG00000116954			-0.84733911

Complex0002	ENSG00000196933	ENSG00000124610
ENSG00000196933	-0.368386636	-4.642569001
ENSG00000124610		-0.637726984

Complex0003	ENSG00000108773	ENSG00000036549	ENSG00000171148	ENSG00000149474	ENSG00000114166	ENSG00000163872	ENSG00000276234
ENSG00000108773	-0.61775377	-4.544818984	-4.357402335	-4.626041398	-4.539036886	-4.333419872	-3.507991897
ENSG00000036549		-0.71716935	-4.548360008	-4.515969075	-4.747723326	-4.688234347	-4.43268168
ENSG00000171148			-0.64245205	-4.930612417	-3.517272045	-3.8203311	-4.52856626
ENSG00000149474				-0.731880193	-5.209156982	-4.547608012	-4.59308534
ENSG00000114166					-0.772059585	-4.256483517	-4.301960846
ENSG00000163872						-0.691234262	-5.145561854
ENSG00000276234							-0.674781529

Complex0004	ENSG00000096746	ENSG00000139675
ENSG00000096746	-0.697390037	-4.527763569
ENSG00000139675		-0.753961137

Complex0005	ENSG00000211456	ENSG00000184432	ENSG00000158623
ENSG00000211456	-0.708655223	-4.670647307	-4.359580051
ENSG00000184432		-0.756757063	-4.668939066
ENSG00000158623			-0.590732441

Figure 7. Partial table of interaction strength: The number corresponding to each pair of genes is the strength of its interaction extracted from the Hi-C data. More negative numbers indicate weaker interaction strengths.

3.5 Separating Table of Interaction Strength into Two Groups

According to the observation of Figure 5., we concluded that the number of marked as "I" tags is 55,113 and the number of marked as "non-I" tags is 3,362. In this case, we filtered out all the gene pairs which have been marked as "I" tags, and calculated the interaction strength between the gene pairs individually, we excluded all data marked as NaN and -inf, and then compared to a distribution for all the rest of the data. Similarly, we have performed a distribution for all gene pairs marked as "non-I" (Figure 8b, c.). These two sets of distributions are very different, so we think this phenomenon is very important for explaining the folding of the 3-D space of chromosomes. In this case, we have categorized the gene pairs into two groups for the further research phases, which are gene pairs located on the same chromosome and gene pairs located on the different chromosomes separately.

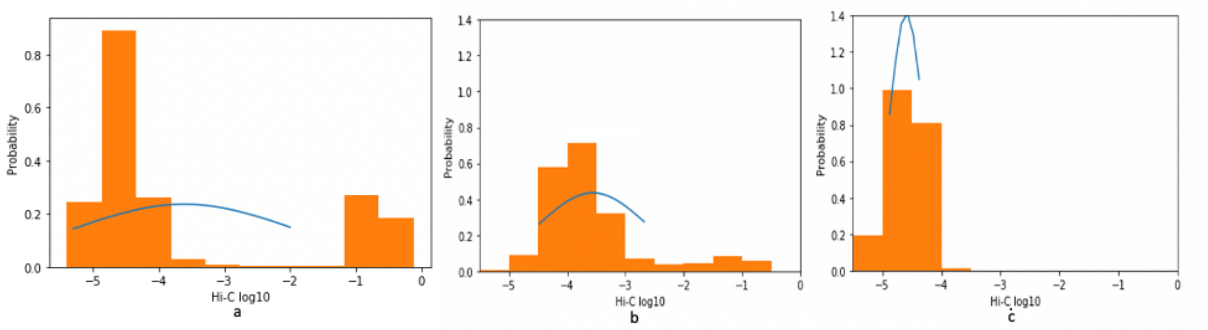

Figure 8. Distributions of interaction strength for all gene pairs and Gene Pairs on the same (b) or different chromosome (c)： a) The distribution for all gene pairs is polarized, with interaction strengths mostly between -5.5 and -4.25. The distribution of interaction strength for gene pairs on the same chromosome (b) differs that of gene pairs on the different chromosomes (c).

3.6 Testing Interaction Strength of Gene Pairs Located on the Same Chromosome

After dividing all real data set into two groups, one with genes located on same chromosome, the other with genes located on the different chromosomes, we focused on testing interaction

strength of gene pairs located on same chromosome and compare real data set with random data set on the same chromosome in this section.

Firstly, we have recollected all the gene pairs located on the same chromosome, and we have calculated the distance difference of the starting positions on the gene pairs. Then we calculated the interact strength between each gene pair, excluded all the data with NaN, and used the reciprocal value to calculate a distribution (Figure 9a.). Then we randomly generated 100 gene pairs at the same distance apart as the real pair. For each gene pair, we repeated the same protocol; finally, the distribution was calculated (Figure 9b.). We observed that the distribution based on the real data, and the randomized data differed slightly, so we have considered the TAD regions. Many studies have shown that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD, therefore we considered all the TAD distribution of the chromosomes (Samantha et al., 2019). We divided the gene pairs located on the chromosomes inside the human protein complex into two groups; the ones that which located on the TAD and not located on the TAD. For each group we then calculated the interaction strength, repeated the previous protocol, and created distribution for each group(Figure 10a. & 10b.). We observed that these two groups' normal distributions differ obviously. In order to test the relationship between the interaction strength and TAD. We hypothesized that if gene pairs are not within a TAD, as long as the distance between them is close enough, they will have strong interaction strength. Throughout our observation, a TAD always extends across at least 80,000 base pairs. So as to the gene pairs which are not located on TAD, we defined the gene pairs whose distance is less than 80,000 as the gene pairs with a shorter distance, the rest of the gene pairs were defined as the gene pairs with longer distance.

Then we performed distribution (Figure 10c. & 10d.) for these two groups' gene pairs based on the above-stated methods. As to the gene pairs with longer distance, we randomly generated 50 pairs of gene pairs with same-length distance and performed distribution (Figure 10e.). These results suggest that gene pairs within TADs have strong interaction strength. And for gene pairs that not in TADs with smaller distance, they still have strong interaction strength, even stronger than gene pairs within TADs.

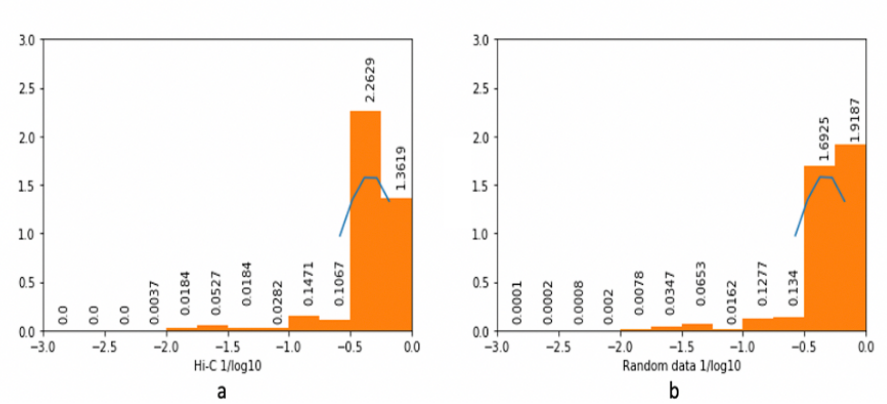

Figure 9 The histogram and distribution of real data and random data (using reciprocal) on same chromosome. a) The histogram and distribution of complex data on the same chromosome: All these figures were originated from normalizations, the value on the top of each bin can represent the probability in each range. **b)** The histogram and distribution of random data on the same chromosome.

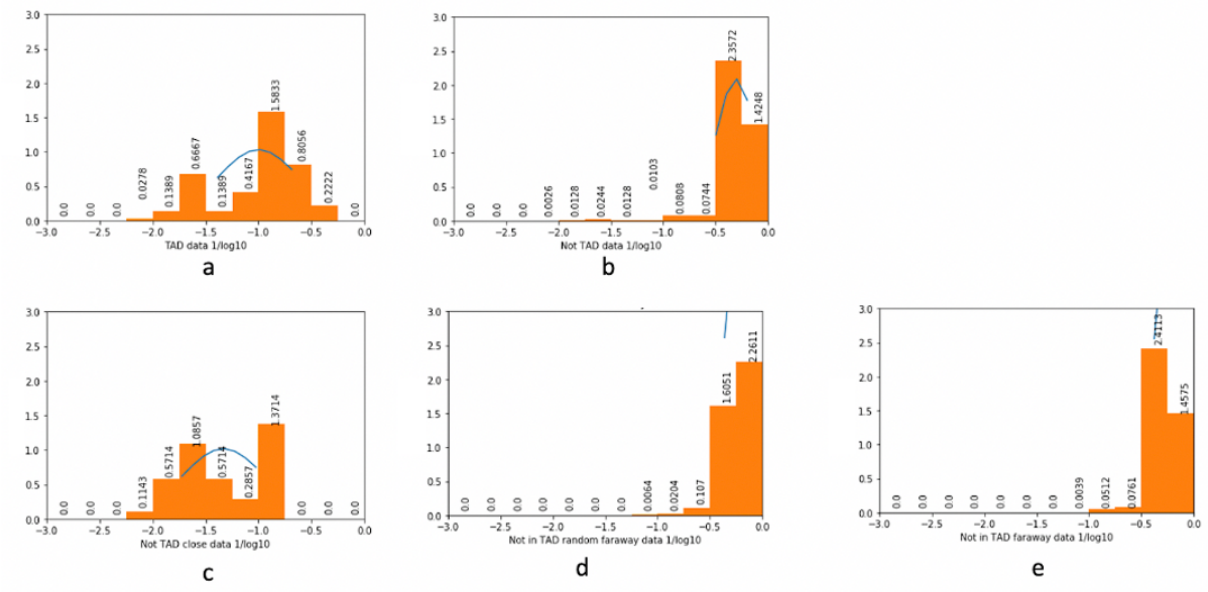

Figure 10 The histogram and distribution of complex random data in or not in TAD region. a) The histogram and distribution of complex data in TAD region. **b)** The histogram and distribution of complex data not in TAD region. **c)** The histogram and distribution of complex data not in TAD region and their distance less than 80,000. **d)** The histogram and distribution of complex data not in TAD region and their distance more than 80,000. **e)** The histogram and distribution of random data not in TAD region and their distance more than 80,000.

3.7 Testing Interaction Strength of Gene Pairs Located on Different Chromosomes

For the gene pairs located on different chromosomes, we still adopted the method of normal distribution analysis (Figure 11a.). Moreover, the research method is the same as the gene pair on the same chromosome. We randomly generated two sets of starting positions on different chromosomes and used the same method for normal distribution (Figure 11c.). If the interaction strengths were related to the fact that the gene pairs encode proteins in a complex, then we would expect to see the distributions of two random datasets would be left tailed distribution. Instead, we see the distributions would be right tailed like the real gene pairs distribution. Through our observations, the two normal distributions are very similar, and it does not explain the conjecture that different chromosomes fold according to the interaction of proteins encoded in their chromosomes. Therefore, in order to test whether the interaction strength on different

chromosomes, based on the two different chromosomes where each gene pair is located in the real data, we randomly generated 100 additional pairs for each group, for a total of 5,511,300 gene pairs. Similar to the above method, we performed a normal distribution analysis on randomly generated gene pairs (Figure 11b.). Our results suggest that the distribution of the three datasets are similar. The results suggest that the genes encoding proteins that have interaction wouldn't in closer physical proximity than the genes encoding proteins that don't have interaction.

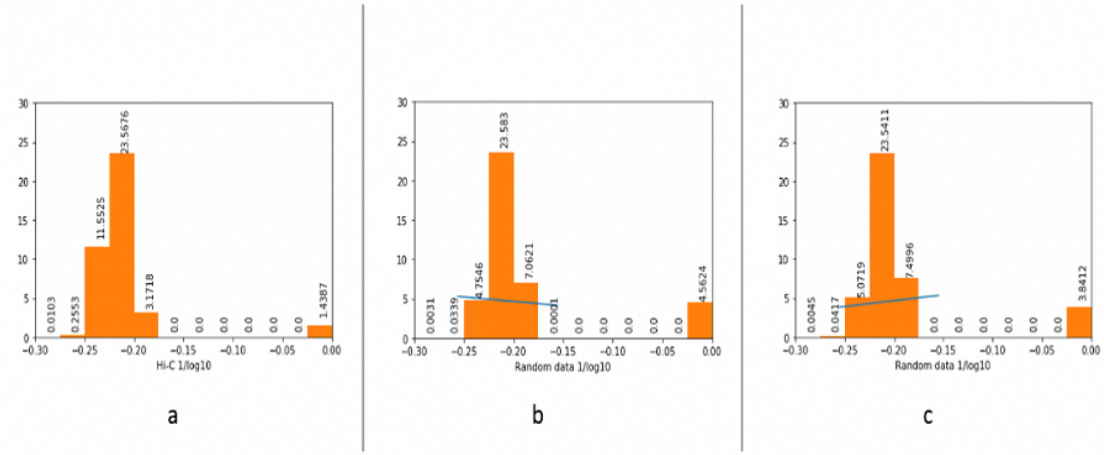

Figure 11 The histogram and distribution of complex and random data on different chromosomes. a) The histogram and distribution of complex data on different chromosomes. **b)** The histogram and distribution of random data on different chromosomes: The random data's chromosome based on the complex data, the chromosome of these randomly generated gene pairs located on is the same as complex data. **c)** The histogram and distribution of random data on different chromosomes: The chromosome of these randomly generated gene pairs located on are randomly generated.

4. DISCUSSION

4.1 Overview of Results

In this study, we examined all the genes encoding proteins known to be found in the human protein complexes. In the distance table between the gene pairs in each complex created in the first step of our analysis, we observed that most of the interactions between genes still come from different chromosomes. We were pleasantly surprised to find that the interaction strength between gene pairs showed a clear polarization when we sorted out the interaction intensity table of all gene pairs. Our hypothesis at this time is that although different chromosomes may fold in the 3-D space, the interaction strength between genes on the same chromosome is still significantly higher. Therefore, in the subsequent analysis, we divided gene pairs into two groups, on the same chromosome or not. After grouping the gene pairs, we first analyzed the distribution of the two sets of data. From Figure 10 a and b, we can see that the interaction intensity distribution of gene pairs located on the same chromosome ranges from -5 to 0, but the interaction intensity of gene pairs located on different chromosomes is all concentrated between -5.5 and -4. This result supports our previous conjecture. Although multiple chromosomes are folded in space, the interactions within a chromosome should have priority.

For these gene pairs located on the same chromosome, we divided the two groups of real data and random data. If the distributions of these two groups were exactly the same, then there would be no difference in interaction strength between genes encoding proteins in complexes and random genes. It can be seen from the analysis that real data genes have the most concentrated interaction strength distribution between -0.5 and -0.25, while random data genes have the most intense interaction strength distribution between -0.25 and 0. But the distribution of the two is

roughly the same, which means that the strength of the interaction between genes encoding these human protein complexes is not as high as we had expected. Therefore, we consider that there will be the influence of TAD on the same chromosome. Within TADs, gene pairs would have much stronger interaction strength, so we want to use the interaction strength of gene pairs within TADs as a threshold to compare them with the gene pairs encoding proteins in human protein complexes. Therefore, we decided to divide these gene pairs encoding human protein complexes and located on the same chromosome into three groups: 1. Very close but not in the TAD. 2. Located in a TAD. 3. Very far away and not located in a TAD. First, in figure 12, a and b proved that the interaction between gene pairs located in TAD is significantly stronger than those not located in a TAD. We were surprised that in the distribution of gene pairs that are very close and not located in a TAD, we observed that their interaction strength is even higher than some gene pairs located in a TAD. The d and e diagrams show that the interaction between gene pairs that are not located in a TAD and are far apart is significantly reduced. In summary, our data show that when located on the same chromosome, TAD is still the first factor to prove the strength of gene interaction, and the distance between genes also has a strong influence. However, based on the small but noticeable difference of distributions of real data and random data, the gene encoding the interacting protein has an effect on the interaction between the genes, but it is not obvious, and the specific reason needs to be proved by further experiments.

For those gene pairs located on different chromosomes, we took a similar approach for analysis. Three sets of distributions were made, corresponding to real data, and two sets of randomly generated control data. In these three sets of distributions, we can intuitively see that no matter whether the chromosomes of random data correspond to the real data, there is no effect or

difference on the interaction strength between the entire genes. Observing the distribution of gene pairs encoding human protein complexes, we can observe that although the distributions are similar among those three figures, the number in the interval -0.25 to -0.2 is significantly higher than that of the other two randomly generated data. This result suggests that we cannot completely rule out the hypothesis we pointed out before that the genes encoding interacting proteins will be located in close physical proximity in 3-D space.

4.2 Limitations of the Experiment and Future Work

For our study, we selected the 4DNFIT3ABRQL dataset in the 4DN data. In all the analyses performed, we chose 10,000 loci as the resolution. Therefore, we have reason to suspect that the current conclusion is not obvious. Since the distance between genes' starting position and ending position may smaller than the resolution, when we compared the gene pairs interaction strengths, we used 10,000 loci that contain the between genes' starting position and ending position rather than precise distance, which may cause error. In the follow-up work, we will try to use a smaller resolution as the smallest unit for analysis. Of course, it will take longer to run all the code. With the development of Hi-C and other technologies and the reduction of high-throughput sequencing costs, the amount of data for genome-wide interactions has increased rapidly, and the resolution of interaction maps has continued to increase. This brings opportunities for the development of 3D genomics but also challenges for computational modeling. At present, the 3D genomic data analysis methods cover a wide range, including data pre-processing, standardization, visualization, feature extraction, and 3D modeling. However, how to choose efficient and accurate methods from them has also become a key factor for our research. But

based on the results we have obtained so far, we remain optimistic about getting more accurate and meaningful conclusions.

References:

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., ... & Marcotte, E. M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, *13*(6).

Oluwadare, O., Highsmith, M., & Cheng, J. (2019). An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological procedures online*, *21*(1), 7.

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., ... & Babu, M. (2012). A census of human soluble protein complexes. *Cell*, *150*(5), 1068-1081.

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B., & Marcotte, E. M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, *13*(6), 932. https://doi.org/10.15252/msb.20167490

Hakim O, Misteli T (March 2012). "SnapShot: Chromosome confirmation capture". *Cell*. **148** (5): 1068.e1–2. doi:10.1016/j.cell.2012.02.019.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of visualized experiments : JoVE*, (39), 1869. https://doi.org/10.3791/1869

Collins, S. R., Miller, K. M., Maas, N. L., Roguev, A., Fillingham, J., Chu, C. S., ... & Ding, H. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature, 446(7137), 806-810.

Zhang, Z., Li, G., Toh, K. C., & Sung, W. K. (2013). 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology*, *20*(11), 831-846.

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., ... & Fraser, P. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome biology*, *17*(1), 127.

Zhang, Z., Li, G., Toh, K. C., & Sung, W. K. (2013). 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology*, *20*(11), 831-846.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., ... & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, *9*(10), 999.

Carstens, S., Nilges, M., & Habeck, M. (2016). Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS computational biology*, *12*(12), e1005292

Adhikari, B., Trieu, T., & Cheng, J. (2016). Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC genomics*, *17*(1), 886.

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., ... & Cramard, J. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, *544*(7648), 59-64

Hung, A. Y., & Sheng, M. (2002). PDZ domains: structural modules for protein complex assembly. *Journal of Biological Chemistry*, *277*(8), 5699-5702.

Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., ... & McKillip, E. (2011). A protein complex network of Drosophila melanogaster. *Cell*, *147*(3), 690-703.

Mukherjee, S., & Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, *19*(7), 955-966.

Branden, C. I., & Tooze, J. (2012). *Introduction to protein structure*. Garland Science.

Schulz, G. E., & Schirmer, R. H. (2013). *Principles of protein structure*. Springer Science & Business Media.

Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. New Science Press.

Fasman, G. D. (Ed.). (2012). *Prediction of protein structure and the principles of protein conformation*. Springer Science & Business Media.

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human mutation*, *17*(4), 263-270.

Nicodemi, M., & Pombo, A. (2014). Models of chromosome structure. *Current opinion in cell biology*, *28*, 90-95.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., ... & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59-64.

Gilbert, D. M., Takebayashi, S. I., Ryba, T., Lu, J., Pope, B. D., Wilson, K. A., & Hiratani, I. (2010, January). Space and time in the nucleus developmental control of replication timing and chromosome architecture. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 75, pp. 143-153). Cold Spring Harbor Laboratory Press.

Adhikari, B., Trieu, T., & Cheng, J. (2016). Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC genomics*, *17*(1), 886.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., ... & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, *9*(10), 999.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, *13*(11), 919-922

Ay, F., Bailey, T. L., & Noble, W. S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research*, *24*(6), 999-1011.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., ... & Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, *9*(10), 999.

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., & Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. *Nature methods*, *14*(7), 679.

Zou, C., Zhang, Y., & Ouyang, Z. (2016). HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome biology*, *17*(1), 40.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268-276.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., ... & Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, *503*(7475), 290-294.

Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, *43*(11), 1059.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., ... & Diao, Y. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331-336

Ozsolak, F., Song, J. S., Liu, X. S., & Fisher, D. E. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nature biotechnology*, *25*(2), 244-248.

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, *14*(6), 390-403.

Supplementary Materials:

All codes, relevant data, and results of this study have been uploaded to:
https://github.com/Jaylen0401/Graduate-thesis

convert.py: This code is used to convert all NCBI Entrez Gene ID from raw data to Ensembl ID.

location.py: This code is used to retrieve relevant information such as the chromosome, start site, and end site where each gene is located by searching Ensembl id.

chart.py: This code is used to map all the genes in the complex one by one to get a table about the distance between the gene pairs.

The following codes were created after connecting to the 4DN server in JupyterHub. Therefore, it is stored in pdf format, and each pdf contains code and result.

Hi-C.pdf: In this code, we combine complex data with Hi-C data to create a table in the same format as chart.py. The result of this code is the interaction strength between the genes in the complex.

test.pdf: In this code, we divided the gene pairs existing in the complex into two groups, which are located on the same chromosome and different chromosomes. Then the two groups were analyzed for normal distribution.

distributions_same.pdf: In this code, we analyzed the distribution of gene pairs located on the same chromosome. After that, the same analysis is performed on the gene pairs randomly generated at the same distance. And use tad data for in-depth analysis.

distribution_diff.pdf: In this code, we analyzed gene pairs located on different chromosomes for distribution analysis, and randomly generated gene pairs for analysis in two different ways.

geneid.csv: This CSV file is the id of all genes in all human protein complexes. Each row represents one complex. 4659 rows are represented in this file.

location0322.csv: This CSV file contains the location of all human protein complex genes. Some of the data that cannot be retrieved is marked with "not found.". The first column is the complex where the gene is located, the second column is the Ensembl id of the gene, the third column is the chromosome where the gene is located, and the fourth column is the starting position of the gene. All unretrievable and problematic data has been deleted. The scale of this data set is 17619 * 4.

Distance_chart.csv: This CSV file shows the table of the distance between each pair of genes in each complex. “I” stands on different chromosomes. The numbers represent the difference between the start points of two genes located on the same chromosome.

tad.csv: This CSV file shows the start and end positions of all TADs on the human chromosome. Each TAD was represented in one line. The first column is the chromosome where the tad is located, and the second and third columns are the start and stop sites.

Same.csv: This CSV file shows all the gene pairs located on the same chromosome. The data has a total of 3362 rows. The first column is the index, and the second and third columns are the chromosome and the start site where one of the genes is located. The fourth and fifth columns are the chromosome and the starting site where the other gene in the gene pair is located.

diff.csv: This CSV file shows all the gene pairs located on different chromosomes. The meaning of each column is similar to the Same.csv. The data totals 55,113 rows.

intad.csv: This CSV file shows all the gene pairs in TAD. The data totals 151 rows. The first column and the second columns are the index of chromosomes and the start site where one of the genes is located. The third and fourth columns are the index of chromosome and starting site where another gene in the gene pair is located. The fifth column is the distance between the two genes.

outtad.csv: This CSV file shows all the gene pairs that are not located in TAD. This data totals 3211 rows. The meaning of each column is the same as the first four columns of intad.csv.

close.csv: This CSV file shows all the gene pairs that are not in the TAD, and the distance is less than 80,000. There are 71 lines of data in total. The meaning of each column is consistent in outtad.csv.

faraway.csv: This CSV file shows all the gene pairs that are not in the TAD, and the distance is larger than 80,000. There are 3140 lines of data in total. The meaning of each column is consistent in outtad.csv.