Effect of DNA Physical Characteristics to Drosophila Nucleosome Positioning

Zehua Li[1,2], Jihua Feng[1,2,*], Xiaowen Zhou[1,2], Huazheng Yu[1,2] and Jing Chen[1,2]
[1]School of Electrical and Information Technology, Yunnan University of Nationalities, Kunming, 650500
[2]Key Laboratory of national cultural resources, Kunming, Yunnan, 650500
[*]Corresponding author

***Abstract*—This thesis has done some research on Drosophila nucleosome positioning in embryonic period, found in nucleosome positioning is related to many factors, such as DNA, CG, DNA content of bending capacity distortion, DNA flip and so on, found that the size of different influencing factors on nucleosome positioning are not the same, here we have twist, roll, tilt and slide four kinds of factors to carry on the regression analysis. It is found that the degree of twist and tilt are more influenced by the degree of roll and slide.**

Keywords-nucleosome positioning; roll; tilt; slide; twist; regression analysis

I. INTRODUCTION

The combination of DNA and related proteins in eukaryotic cells is a chromatin.[1]. In all eukaryotic organisms, the basic subunit of chromatin is formed by octamer which has two copies of four core porteins(H2A, H2B, H3, and H4). DNA is winding nucleosome structure of the formation of the "bead like" in the group of protein particles.

The majority of DNA in eukaryotic cells are packed by the nucleosomes, each of which has a roughly 147 bp of DNA. But whether in the interphase nuclei of euchromatin or heterochromatin, or in mitotic chromosomes, which are the constant component of the nucleosome.

The exact location of the DNA on the genome is called the nucleosome position. The positioning of the nucleosome bodies can be realized in any of the following two ways: 1, intrinsic: each of the nucleosome bodies is specifically placed on a specific DNA sequence. This corrected our view that the formation of the sequence as a subunit, which can be composed of any DNA sequence and the histone octamer, is not correct. 2, external: the first in a certain region of the specific location of the assembly will be assembled in the special positioning point. Because there is no nucleosome region, so nucleosome positioning have a preference for the starting point, this area provides a limit for the adjacent nucleosome available sites, and then a series of nucleosomes with a certain repeat length according to the order of assembly.

For the study of the positioning of the nucleosome, we can think like this, with different biological cell structure, its chromatin and chromatin in the DNA structure is also different. The genome of single cell organisms such as yeast is small, and the characteristics of the positioning of their nucleosome are relatively simple, so what are the characteristics of the nucleosome position of multicellular eukaryotes? In order to study the chromatin structure of multicellular eukaryotes more convenient, here we chose nucleosome positioning data of Drosophila embryos. The reasons are as follows: 1) Drosophila embryo from a single cell to differentiation of multicellular stage, the chromatin structure of this period, compared with the direct study of differentiation the adult flies are much smaller. 2) Compared with the single cells of yeast cells, Drosophila embryo possesses some characteristics of unicellular organisms, thus obtained in the study of qualitative staining in yeast, can be used to study the Drosophila, two were comparable; 3) Predecessors have been obtained nucleosome positioning data of Drosophila embryonic stage[4] through experiments, so the research has feasibility.

II. EXPERIMENTAL METHOD

A. Data Preparation

Based on the data from the two main parts: first, the yeast nucleosome positioning experiment data, including high resolution nucleosome occupancy of the experimental data in the study by Lee et al[5], 16 yeast chromosomes encoding sequence of DNA from the NCBI database, 4792 high confidence yeast genome data which presented by David et al in the literature, and H2A.Z nucleosome position data got by Albert et al through experiment. The second part of the experimental data of nucleosome positioning in Drosophila, including the the Drosophila embryo H2AZ nucleosome positioning data (bulk nucleosome) got by Mavrich et al [2], the overall nucleosome occupancy data, and classification data on Drosophila embryo gene expression pattern got by Pavel Tomancak et al[6]. Since the above nucleosome position experimental data is from the different experimental platform, we accord to the research purpose to use the signal processing method to reconstruct the data.

By preliminary treatment and screening, the position data of the whole genome of Drosophila (all genes) were obtained. This set of data includes the location of genes on the chromosome, the initiation and termination sites of the gene, the positive transcription (W) and reverse transcription (C) of the gene.

Figure I lists the Drosophila gene data, of which second columns of gene name, the third column indicates the number of genes that are located on the chromosome. Fourth columns represent the direction of transcription. Sixth columns and fifth columns show transcription start sites and transcription termination sites. When the fourth column is 'W' ('+'), it is

positive transcription, transcription direction of transcription is initiation site to transcription termination sites; however, when the fourth column is 'C' ('-'), this time is reverse transcription, namely, the direction of transcription is transcription termination sites to transcription initiation site.

	1	2	3	4	5	6
1	'FBtr01005...	'CG6741'	2	'W'	18024494	18060339
2	'FBtr00802...	'CG4807'	1	'W'	11210681	11261081
3	'FBtr00833...	'CG10325'	4	'C'	12655769	12633344
4	'FBtr00833...	'CG11648'	4	'C'	12797958	12752932
5	'FBtr00753...	'CG4032'	3	'C'	16641674	16615461
6	'FBtr00801...	'CG6093'	1	'C'	10975273	10973442
7	'FBtr00700...	'CG3796'	6	'W'	264064	265024
8	'FBtr00827...	'CG17907'	4	'C'	9084590	9053962
9	'FBtr00740...	'CG9151'	6	'C'	15272487	15246390
10	'FBtr00856...	'CG7899'	4	'C'	25819130	25816863
11	'FBtr00846...	'CG5610'	4	'C'	20282651	20226948
12	'FBtr00723...	'CG4356'	2	'C'	20277235	20266159
13	'FBtr00732...	'CG11348'	3	'W'	4429370	4435999
14	'FBtr00846...	'CG6844'	4	'W'	20311164	20316875
15	'FBtr00708...	'CG4027'	6	'W'	5794897	5798210
16	'FBtr00860...	'CG12051'	2	'C'	1903190	1901415
17	'FBtr00715...	'CG10067'	2	'W'	16831533	16833945
18	'FBtr00784...	'CG7478'	3	'W'	21978331	21980290
19	'FBtr00827...	'CG18290'	4	'W'	9251707	9253810

FIGURE I. THE TRANSCRIPTIONAL DATA OF THE DROSOPHILA

B. Data Processing

1) Processing of experimental data of high resolution rate of occupation of the core: Because of uneven data, therefore, here we obtained data by interpolation method is: covering the whole genome of each of the 1 bp nucleosome occupancy data, these data for subsequent alignment using. And then through the three spline interpolation (spline), we obtained the results of the 6 chromosomes of the fruit fly with higher accuracy and better smoothness. Due to the large interval (36 bp) of the Drosophila data, the accuracy of the experimental data obtained from the 6 chromosomes of the fruit fly is relatively low.

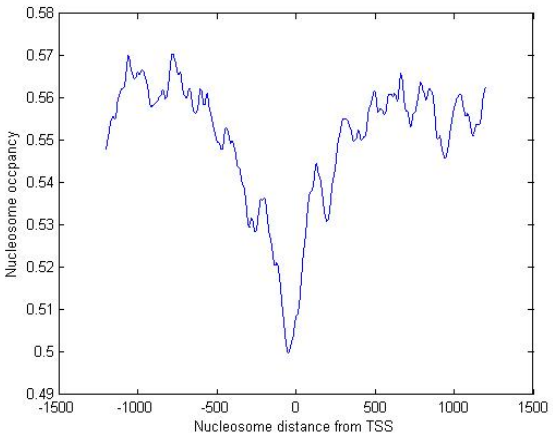

FIGURE II. THE MAP OF DROSOPHILA NUCLEOSOME POSITIONING

2) Data sorting based on center alignment: In order to get the map of each group of genes. The arrangement of each gene in accordance with the principle of increasing the length of the gene from small to large order respectively, and extract the center point of each gene, so we get each position of gene from short gene to long gene . Secondly, this position data were mapped to the whole data of the whole space occupying rate of the fruit fly and the positioning data of H2A.Z. Find the center point of each gene on each chromosome, and around the center point of each gene, the length range of the 1200bp is intercepted. Finally, align the intercepted data(type C gene corresponding data to be in reverse processing), so as to get the alignment of each gene center of the positioning data. Drawing the aligned data, the map of the alignment of the gene centers in each group was obtained. As shown in Figure II.

A. Lasso Regression Algorithm

Lasso algorithm[8] is the selection and shrinkage operator least absolute. In linear regression, the L2 model is usually used as a penalty function. Although the L2 paradigm is effective and stable in dealing with some aspects of the problem, but the dynamic range is small, especially when it comes to the need to make a detailed resolution of the model coefficients, the L2 paradigm has a congenital deficiency. In view of this kind of question, if using the L1 model, the reduction factor can be more effective, so that the model is more easy to explain. The Lasso algorithm for unconstrained condition is defined as follows:

$$\|Xw - y\|_2^2 + \lambda\|w\|_1 \tag{1}$$

Obviously, this is a convex optimization problem based on the weight. But at the time when w_i=0, the formula (1) can not seek derivative. Therefore, a closed analytic expression of the global minimum value, which is similar to the L2 paradigm, can not be obtained. Taking into account the specific the problem of nucleosome positioning, this paper selects a representative of the iterative ridge regression algorithm[7] to solve the problem. The method is similar to the Newton method:

$$|w_i|_1 \approx \frac{w_1^2}{|w_i|_1} \tag{2}$$

The type (2) into the equation (1), and can get a similar to the L2 paradigm as the least squares solution of the penalty function:

$$w_{new} = X^T X + \lambda \mathrm{diag}(|w|)^{-1} X^T y \tag{3}$$

B. Regression Analysis

In order to explain the possible influence factors on Drosophila nucleosome positioning, this paper puts forward the following simple hypothesis: (1) the interaction between factors on the formation of Drosophila nucleosome position is a linear additive. (2) in addition to the space position, the influence of the orientation between adjacent bodies is independent of each other. On the basis of this, we trained a linear regression model, fitting the multiple DNA related

location factors with the position data of the gene on the expression gene of Drosophila embryonic period.

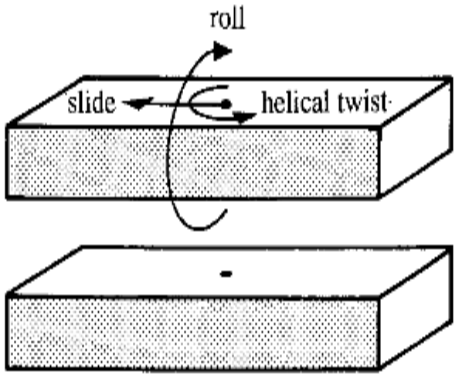

FIGURE III. THE RELATIVE POSITION RELATIONSHIP BETWEEN SOME DNA BASE STEPS[3]

The map shows the possible formation of the two adjacent bases in the DNA deformation, which are: roll, twist and slide.

In several regression algorithms, Lasso algorithm can effectively overcome the phenomenon of over fitting. Therefore, we use Lasso regression to build a linear model, as follows:

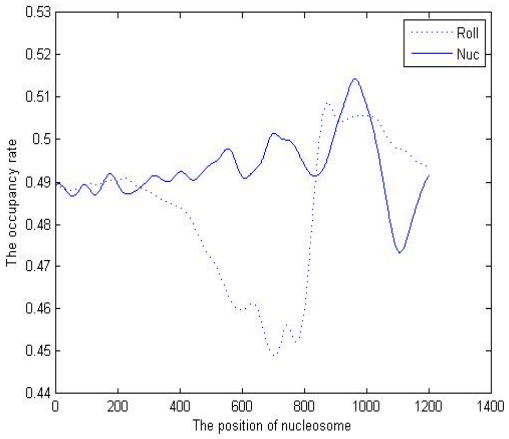

FIGURE IV. THE CONTRAST BETWEEN ROLL AND ALL NUCLEOSOME POSITIONING

Pictured above, is established by using Lasso algorithm by a factor of Roll data fitting and comparative map of nucleosome positioning map, we set the transcription initiation site in 1000bp, comparing two curves, the difference is relatively obvious between Roll curve and Nuc curve (here refers to all the nucleosome positioning curve). Roll curve change more between 400-800bp but Nuc curve change less. Nuc curve change more between 800-1200bp but the Roll curve stay in a small fluctuation. The fluctuations in the 0-400bp; the two curves of the difference is not great, indicating that Roll has effect on the nucleosome positioning, but not much.

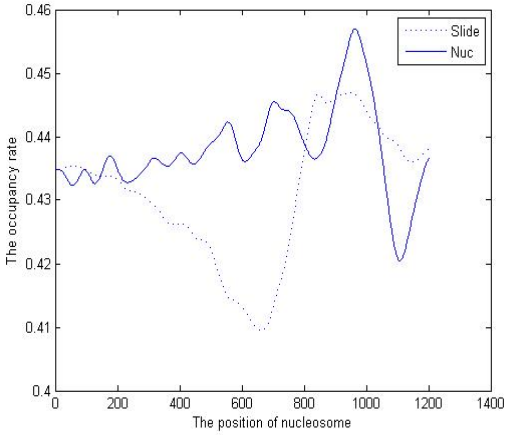

FIGURE V. THE CONTRAST BETWEEN SLIDE AND ALL NUCLEOSOME POSITIONING

Pictured above, is established by using Lasso algorithm by a factor of Slide data fitting and comparative map of nucleosome positioning map, we set the transcription initiation site in 1000bp, comparing two curves, the difference is relatively obvious between Slide curve and Nuc curve (here refers to all the nucleosome positioning curve). Slide curve change more between 200-800bp but Nuc curve change less. Nuc curve change more between 800-1200bp but the Slide curve stay in a small fluctuation. The fluctuations in the 0-200bp; the two curves of the difference is not great, indicating that Slide has effect on the nucleosome positioning, but also not much like the Roll.

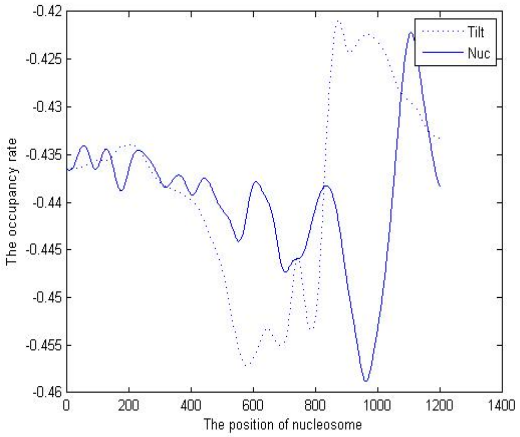

FIGURE VI. THE CONTRAST BETWEEN TILT AND ALL NUCLEOSOME POSITIONING

Pictured above, is established by using Lasso algorithm by a factor of Tilt data fitting and comparative map of nucleosome positioning map, we set the transcription initiation site in 1000bp, comparing two curves. Between 0-400bp, Tilt curve is relatively stable, and relatively sharp in 400-1200bp. Between 0-400bp, the Nuc curve of the same fluctuations is small, and relatively sharp fluctuations between 400-1200bp. Relative to the previous two factors (Roll, Slide), the effect of Tilt on the nucleosome positioning is bigger. And the general

trend of the two curves is the same, here we think that the DNA base pair tilt (Tilt) has a relatively large influence on the positioning of the nucleosome.

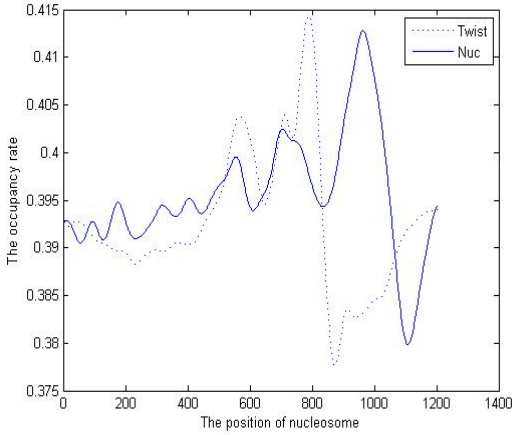

FIGURE VII. THE CONTRAST BETWEEN TWIST AND ALL NUCLEOSOME POSITIONING

Pictured above, is established by using Lasso algorithm by a factor of Twist data fitting and comparative map of nucleosome positioning map, we set the transcription initiation site in 1000bp, comparing two curves between 0-700bp Twist curve is relatively stable, relatively sharp in 700-1200bp. Between 0-700bp, the Nuc curve of the same fluctuations is small, relatively sharp fluctuations between 700-1200bp. Relative to the previous two factors (Roll, Slide), the effect of Twist on the nucleosome positioning is bigger. And the general trend of the two curves is the same, here we think that the twist degree has a relatively large influence on the positioning of the nucleosome.

III. CONCLUSION

Although the Drosophila as a well-known genetic model organism has a history of nearly one hundred years, but at the molecular level of the chromatin structure of the nucleosome unit is just beginning. The first problem encountered in the study is the available experimental data far from the abundance of single cell organisms. In this paper, the only group of Drosophila embryonic stage of the positioning data were analyzed, with introduction of some new conclusions, we also pointed out the new problems that we encountered. Some of these problems have been resolved, and some are only partially resolved, and some problems have yet to be clarified or confirmed by the appearance of new materials.

In order to evaluate the influence of DNA physical properties on the orientation of the nucleosome, this paper uses the Lasso regression algorithm to predict the position of the drosophila. According to the interpretation of the diagram, we found that the DNA parameters in the degree of distortion (Twist) and the inclination (Tilt) have a greater impact. Description in the Drosophila genome, to the large extent, the packaging of DNA to the nucleosome could be restricted by consumption of energy required for deformation[8].

This paper only selects and analysis four factors to the influence of nucleosome positioning, and then we will choose other factors, and analyzes their impact on nucleosome positioning.

REFERENCES

[1] Cai L., and Zhao X.L., 2009, Advances in nucleosome positioning, Shengwu Wulixue Bao (Acta biophysica sinica), 25(6): 384-395.

[2] Mavrich TN, Jiang C, Ioshikhes IP. Nucleosome organization in the Drosophila genome[J].Nature,2008 May 15;453(7193):358-62.

[3] Hassan MA, Calladine CR. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA[J]. J Mol Biol, 1996 May 31;259(1):95-103.

[4] Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol, 2007, 8(7): R145.

[5] Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet, 2007, 39: 1235-1244.

[6] Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol, 2007, 8(7): R145.

[7] Neumann A, Holstein J, Chatellier G. A regression shrinkage method tailored to qualitative regressors and clustered data[J]. Stat Med, 2004 Apr 15;23(7):1147-57.

[8] Feng J, Dai X., Xiang Q.,Dai Z., Wang J., Deng Y., He C., 2010,New insights into two distinct nucleosome distributions: comparison of cross-platform positioning datasets in the yeast genome.BMC Genomics, 77.