Partial Study on the Genesis of Nucleosome Localization in Drosophila

Zehua Li, Yongfa Ling* and Xiaoyu Li
Hezhou University, Hezhou, China
*Corresponding author e-mail: yfling73@163.com, b523788254@163.com, xiaoyu1846916@163.com

Abstract. This paper has done some research on the nucleus localization of Drosophila in embryonic stage. It is found that the influence of different factors on the location of nucleosomes is not the same. It includes not only the physical properties of DNA, such as DNA twist, roll, tilt, and slide, but also other factors, such as DNA bending ability, CG content, transcription factors, RNA polymerase, and others. Here we carry out regression analysis on these factors, we finally get the weight of these factors in the nucleosome positioning process fitting by LASSO algorithm.

1. Introduction
In recent years, with the development of science and technology, the results of biological research have been applied to various fields, and the research of genomics has also been greatly developed. Scholars have gradually focused their attention on the exploration of various biological DNA. As a result, the field of life sciences has ushered in a new era of development. A huge amount of data is constantly being produced, discovered, and believed to continue to grow rapidly in the future. But such huge data gives us more confusion about the exploration of life than surprises. How to interpret these data will become a huge challenge for bioinformatics research for a period of time [1].

Whether in prokaryotes or eukaryotes or even viruses, the structure of DNA is constant. Further, DNA interacts with histones to form a chromosome. In all eukaryotes, the composition of the basic subunit of chromatin is also invariant, and all are octamers with four histones as the core. These granular octamers are entangled on DNA like beads, and this bead-like structure, which is formed by DNA entanglement on the surface of histone particles, is called nucleosome [2].

In recent years, as the nucleosome localization data of the Drosophila embryonic stage has gradually improved, it has provided more convenient conditions for the researchers to further study the chromosome organization and structure of eukaryotes. When studying the nucleosomes of many eukaryotes, there are several reasons for choosing fruit flies as research objects. Firstly, cultivation is relatively convenient, because the fruit fly itself is a relatively small species, it takes up less space throughout the cultivation process. Moreover, the requirements for equipment and feeding materials are relatively low during the cultivation process. In a conventional culture vessel, a substrate for fermenting yeast is used as a medium, that is, a large-scale culture can be carried out, which is in accordance with conditions for actual cultivation. Secondly, the relatively high frequency of reproduction. During the whole process of breeding, from egg, larva, pupa to forming process, the temperature requirements are low, reaching room temperature of 25 degrees Celsius, and we also have the basic conditions for the
cultivation of the laboratory, and its life process is short, only 10 days. Thirdly, there is a relatively simple genetic trait. The genome of Drosophila nucleus has only 6 pairs of chromatin, it is easy to identify male and female individuals and unmatured individuals, which have higher recognition in the experiment. Fourthly, the gene is highly conserved. In comparison with other higher animals and human genes, drosophila genes are highly conserved in structure and function, and their research results can be used by other species.

2. Experimental method

2.1. Data Preparation

The data of this study mainly come from two parts, First, the experimental data of yeast nucleosome localization, including the high-resolution nucleosome occupancy rate experimental data given by Lee et al., yeast’s 16 chromosomal DNA coding sequences from the NCBI database, the 4792 high-confidence yeast genome experimental data given in the literature by David et al and the H2A. Z nucleosome position data obtained by Albert et al. The second part of the experimental data of nucleosome localization of Drosophila includes the Drosophila embryonic nucleosome localization data and the bulk nucleosome occupancy rate data obtained by Mavruch et al. Since the above nucleosome localization experimental data comes from different experimental platforms and has heterogeneity, we reconstructed the data related to this study by using signal processing methods according to the research purpose.

2.2. Data processing

We performed the binarization of the experimental data of the genus of the genus drosophila. Set all the nucleosomes on the corresponding DNA sequence to 1, and the place without nucleosomes to 0. This results in binarization data for the corresponding 6 Drosophila chromosomes with 0, 1 representing the presence or absence of nucleosomes.

As shown in Fig. 1, the first column symbol indicates the name of the Drosophila gene. The second column symbol indicates the unique code for the Drosophila gene. The third column of numbers indicates the Drosophila chromosome serial number. 1 represents the first chromosome, 2 represents the second chromosome, 3 represents the third chromosome, 6 represents the sixth chromosome, and so on. The fourth column symbol indicates the type of transcription, and there are two types of W type and C type, which indicate positive transcription and reverse transcription, respectively. The W type represents forward transcription, and the chromosomal transcription is transcribed from the transcription start site to the transcription termination site, that is, from left to right. The C type refers to reverse transcription, and transcription is from the transcription termination site to the transcription start site, that is, from right to left. The fifth column indicates the transcription start position. The sixth column indicates the transcription termination position.

Taking the 14141 gene transcription start site (TSS) in Figure 1 as the center point, the 1200bp nucleocapsid occupancy rate data was selected upstream of the TSS, and the 1200bp nucleocapsid occupancy rate data was also selected downstream of the TSS. These data are aligned, here we consider the directionality of gene transcription, and the nucleosome occupancy rate data corresponding to the C type gene should be mirrored and folded, then superimposed and averaged. In order to facilitate the next research and analysis, we smoothed and normalized the data obtained. Finally, the average occupancy map of nucleosomes centered on the gene TSS is obtained, as shown in Fig. 2.
3. Regression analysis of positioning factors

3.1. Lasso regression algorithm

The Lasso algorithm is an abbreviation for Least Absolute Selection and Shrinkage Operator. In linear regression, the L2 paradigm is usually used as a penalty function. Although the L2 paradigm is effective and stable in dealing with certain aspects, the L2 paradigm has congenital shortcomings when the dynamic range is small, especially when it comes to detailed resolution of model coefficients. For this type of problem, if the L1 paradigm is adopted, the coefficient can be reduced more effectively, making the model easier to explain. The Lasso algorithm for unconstrained conditions is defined as follows.

$$\|Xw - y\|_2^2 + \lambda \|w\|$$

(1)

Obviously, this is a problem of solving convex optimization based on weights. But at the time, $w_i = 0$, equation (1) is not conductive. Therefore, a closed analytical expression that is similar to the global minimum of the L2 paradigm cannot be obtained. Considering the specific nucleosome positioning fitting problem, this paper chooses the representative unconstrained iterated Ridge Regression.
algorithm to solve and get $w$. This method makes an approximation similar to the Newton method, namely,

$$|w_i| \approx \frac{w_i^2}{|w_i|}$$

Substituting equation (2) into equation (1) yields a least squares solution similar to the L2 paradigm as a penalty function:

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

3.2. Regression analysis

In order to explain the extent of the influence of various possible factors on the location of Drosophila nucleosomes, this paper first proposes the following simple hypotheses. First, the relationship between the positioning factors and the formation of Drosophila nucleosomes is linearly additive. Second, in addition to spatial positional constraints, the positioning effects between adjacent nucleosomes are independent of each other. On this basis, we trained a linear regression model, using multiple DNA-related localization factors to fit the experimentally observed nucleosome location data of the Drosophila embryonic expression gene.

Among several regression algorithms, the Lasso algorithm can minimize the weight of the correlation factor, thereby assigning a larger weight to the factor contributing more, and can effectively overcome the over-fitting phenomenon. Therefore, we used Lasso regression to establish a linear model.

Fig. 3 is a comparison map between RNA polymerase II and nucleosome localization curves. According to the transcription principle, we can know that RNA polymerase II plays an indispensable role in the whole gene transcription process as an important catalytic condition. Therefore, the process of polymerase II extension through the gene coding region is likely to cause differences in the distribution of nucleosomes before and after transcription. Therefore, the difference in distribution between nucleosomes before and after transcription is likely due to the action of RNA polymerase. However, it can be seen from the figure that the difference between the factor of RNA polymerase II and the nucleosome localization map is still relatively large, so we still hold the influence of the factor of RNA polymerase II on the location of nucleosomes. There is a little reservation, at least for now, the weight of the nucleosome positioning will not be large.

![Figure 3. Comparison pol2 and nucleosome localization mapping.](image)

Fig. 4 shows a comparison map between transcription factor (TFBS) and nucleosome localization curve. The nucleosome localization is affected by transcription factors, which are specifically affected
by competing with nucleosomes for DNA binding sites. Many transcription factors bind to DNA-specific sites to cause both bending and closure effects on the surface of the target site, which causes the target site to form a completely different conformation to the nucleosome DNA. From the figure, we found that the factor of transcription factor (TFBS) is far from the nucleosome. We can think that the competition of nucleosomes and transcription factors on target sites will affect nucleosome localization. The degree of affinity between nucleosomes and DNA and between transcription factors and DNA determines the degree of competition. At the same time, their concentration in the same region will also affect the degree of nuclear competition. The combination of the two factors indirectly affects the location of nucleosomes. We found that transcription factors can only be combined when the genome completely dissociates or removes the nucleosome from its occupied transcription factor binding site, and the entire process requires the interaction of other chromatin remodeling factors. Therefore, this process is not the result of the separate action of transcription factors [3-5]. Therefore, the role of transcription factors in this process is part of the whole, but the analysis is not taken out.

![Figure 4. Comparison of transcription factor (TFBS) and nucleosome localization.](image)

Fig. 4 is a comparison map between the CG content and the nucleosome positioning curve. From the figure we can observe the difference between it and the nucleosome positioning curve, especially between -200-300bp. Although there are dinucleotide sequences in each DNA sequence, that is bases containing AA/TT/TA/AT and CC/CG/GC/GG. At the same time the nucleosome localization is also bound to be affected by the base content. The base content in each chromosome is not fixed, so the amount of these bases will affect the location of the nucleosomes.

![Figure 5. Comparison of CG content and nucleosome localization.](image)

Fig. 5 is a comparison map between the CG content and the nucleosome positioning curve. From the figure we can observe the difference between it and the nucleosome positioning curve, especially between -200-300bp. Although there are dinucleotide sequences in each DNA sequence, that is bases containing AA/TT/TA/AT and CC/CG/GC/GG. At the same time the nucleosome localization is also
bound to be affected by the base content. The base content in each chromosome is not fixed, so the amount of these bases will affect the location of the nucleosomes.

Fig. 6 shows a comparison map between CTG content and nucleosome localization curve. From Figure 6, we can also directly observe the difference between CTG content and nucleosome localization map. For the relationship between CG content and CTG content and nucleosome localization, we have the following conclusions. A dinucleotide sequence in each DNA sequence, like the base pairs of A: A, T: T, T: A, A: T and C: C, C: G, G: C, G: G. Nucleosome localization is also affected by its base content. For example, if the A: T content in a base pair is relatively high, then its DNA will show a tendency to bend toward the minor groove. At this time, the direction of the minor groove is toward the histone octamer, which is advantageous for the assembly of the nucleosome [6, 7]. On the contrary, if the G: C content in the base pair is relatively high, then its DNA will show a tendency to bend back toward the small groove. At this time, the direction of the minor groove is away from the histone octamer, which is not conducive to the assembly of the nucleosome.

![Figure 6](image1.png)

**Figure 6.** Comparison of CTG content and nucleosome localization.

Fig. 7 shows a comparison map between DNA bending ability and nucleosome localization curve. In the composition of nucleosomes, DNA will entangle on the octamer histones, and the entanglement will definitely cause deformation, at the same time the bending ability of DNA is an important factor in the degree of entanglement. We can understand that if the bending ability of DNA is strong, then its ability to wrap is relatively strong. On the contrary, the bending ability of DNA is weak, its flexibility will inevitably decline, so the winding ability is naturally relatively weak. Therefore, DNA bending ability will inevitably affect the composition of nucleosomes.

![Figure 7](image2.png)

**Figure 7.** Comparison of DNA bending ability and nucleosome positioning.
3.3. LASSO fitting analysis

After training the Lasso regression model in the promoter region, we obtained the weight values of each positioning factor in the entire nucleosome positioning, as shown below.

From the figure, we can also visually observe the weight value of each factor in the whole nucleosome positioning process which representing the contribution of these factors to the location of nucleosomes in the region. In the fitting process of the nucleosome localization cause, the LASSO algorithm will also get the result of the difference because of the difference in the region of the chromosome. In actual experiments, it was found that the results of the calculations on the entire chromosome, in the coding region and between the intergenic regions are different, which indicates that the nucleosomes are located in different regions, such as in. The differences between the drosophila gene coding regions and intergenic regions are limited by different mechanisms. In the total nucleosome localization map, it can be seen that the higher weight is always assigned to two factors, namely the DNA twist (Twist) and the DNA tilt (Tilt). Explain that during the nucleosome localization of Drosophila, the factor that has the greatest impact on it is the energy consumed when DNA undergoes deformation during the packaging of nucleosomes. Moreover, the two factors of DNA distortion and inclination, as extremely component parameters, occupy an important position in the periodic process of DNA conformation, so the correlation between the two is highly correlated with the location of the nucleosome in the promoter region. In the above experiments, the two factors of DNA distortion and DNA tilt were opposite in the process of nucleosome localization, the former was positively correlated with nucleosome localization, while the latter was negatively correlated, which is consistent with the point mentioned by Hassan and Calladine in their article entitled "Spiral entanglement of bases in DNA and mobility of dinucleotides.

Figure 8. Regression model impact factor weights.

4. Summary

Although some progress has been made in the above research work. However, there are still many shortcomings in this paper, which need to be improved and improved in the subsequent work. First of all, the factors affecting the location of nucleosomes are multifaceted. Because of the limitations of time and other conditions, the factors that lead us to study are not comprehensive enough, and there will be a lack of analysis of individual factors. Secondly, in this study, only the LASSO algorithm-based study on the cause of Drosophila nucleosome localization was considered, and no other studies were conducted. In the following work, we also need to apply other algorithms in the study of Drosophila nucleosome localization, such as support vector machines, genetic algorithms, etc. Only by comparing the conclusions of these algorithms can we get more convincing results.

Acknowledgments

This research was supported by the National Science Fund of Guangxi, Grant 2018JJA170185,
Compressive Sensing Processing and Application Research of Water Quality Monitoring Data in Wireless Sensor Networks. And “Cognitive Radio and Information Processing” Key Laboratory Fund Project of the Ministry of Education, Research on Key Technologies of Data Compression and Noise Reduction Processing in Wireless Sensor Networks.

References
[1] Rubin G.M., and Lewis E.B., 2000, A brief history of Drosophila's contributions to genomeresearch, Science, 287 (5461): 2216-2218.
[2] Zhang Z, Pugh B F. High-resolution genome-wide mapping of the primary structure of chromatin [J]. Cell, 2011, 144(2): 175-186.
[3] Yu L, Morse R H. Chromatin opening and transactivator potentiation by RAP1 in Saccharomyces cerevisiae [J]. Molecular and Cellular Biology, 1999, 19 (8): 5279-5288.
[4] Adkins M W, Tyler J K. Transcriptional activators are dispensable for transcription in the absence of Spt6-mediated chromatin reassembly of promoter regions [J]. Molecular cell, 2006, 21 (3): 405-416.
[5] Kim H D, O'shea E K. A quantitative model of transcription factor–activated gene expression [J]. Nature structural & molecular biology, 2008, 15 (11): 1192-1198.
[6] Widom J. Role of DNA sequence in nucleosome stability and dynamics [J]. Quarterly reviews of biophysics, 2001, 34 (03): 269-324.
[7] Zhurkin V B, Lysov Y P, Ivanov V I. Anisotropic flexibility of DNA and the nucleosomal structure [J]. Nucleic acids research, 1979, 6 (3): 1081-1096.