Nucleosome Positioning with Set of Key Positions and Nucleosome Affinity

Jia Wang¹, Shuai Liu²,³*, Weina Fu²

¹Experimental Instrument Center, Dalian Polytechnic University, Dalian, Liaoning, 116034, China
²College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia, 010012, China
³School of Physical Science and Technology, Inner Mongolia University, Inner Mongolia, 010012, China

Abstract: The formation and precise positioning of nucleosome in chromatin occupies a very important role of study in life process. Today, many researchers discover that the positioning where the location of a DNA sequence fragment wraps around a histone in genome is not random but regular. However, the positioning is closely relevant to the concrete sequence of core DNA. So in this paper, we analyzed the relation between the affinity and sequence structure of core DNA sequence, and extracted the set of key positions. In these positions, the nucleotide sequences probably occupied mainly action in the binding. First, we simplified and formatted the experimental data by the affinity. Then, to find the key positions in the wrapping, we used neural network to analyze the positive and negative effect of nucleosome generation for every position in core DNA sequences. However, we reached a class of weights with every position to describe this effect. Finally, based on the positions with high weights, we analyzed the reason why the chosen positions are key positions, and used these positions to construct a model of nucleosome positioning predict. Experimental results show the effectiveness of our method.

Keywords: Nucleosome positioning, Affinity, DNA sequence, Key position, Neural network.

1. INTRODUCTION

Admittedly, nucleosome is the basic structure unit of chromatin. A nucleosome is constructed by a DNA fragment (core DNA) and a histone. Usually, the length of DNA around histone is about 147 basepair, and wrapped over a histone about 1.65 circles. Though the core DNA length is different with different organisms, kinds of cell and areas of chromatin, it is known that nucleosome occupies 75% - 90% of genome. It mean that nucleosome plays a role in life process. In fact, researchers find that nucleosome positioning plays a role in transcription regulation, gene expression and splicing [1]. However, not all base-pairs function equally in Histone wrapped. Meantime, different kinds of histone show different preference to DNA fragment. It means that nucleosome has its DNA sequence preference [2-4].

Earlier, Komberg first presented nucleosome positioning based on statistics with barrier model [5]. He found that the nucleosome positioning is highly certain. In recent years, the statistical model was under highly research. Yuan and Mavrich etc. researched in the statistical model and found the nucleosome positioning obeyed statistics outside barrier because of the electrostatic and steric hindrance effect [6-8]. They found that determinacy of nucleosome is lower when the position is farther from the barrier. Then, Fu and Schones studied in nucleosome positioning by human genome, and supported viewpoint of statistical positioning by the analysis for difference of yeast and human [9-10]. Zhang and Stein also found that the DNA sequence preference determines mainly in rotating position of nucleosome, but limited in translational displacements [11, 12]. Ioshikhes counted and computed standard distribution of AA/TT in core DNA sequence [13]. Leimgruber further compared distributional correlation of diad AA/TT both in experimental DNA sequence and standard distribution. He found that lack of nucleosome is corresponding with the valley region of associated curve, and center of nucleosome is corresponding with the peak region of associated curve [14].

In recent years, there were more predicting models presented with the appearance of in vivo nucleosome positioning data sample [15, 16]. Zhao etc. classified nucleosome preference sequence and repellence sequence of yeast, drosophila and human by applied diversity of increment [17]. Their high level accuracy supported the viewpoint of DNA sequence positioning. Liu etc. use curvature profile model to predict properties of nucleosome positioning at target site of TSS, TFBS, SNP and miRNA [18]. Recently, Becker etc. presented a variable optimal statistical model in nucleosome positioning [19]. This model use a study-predict method to predict probability distribution of nucleosome.

Since Segal present his viewpoint that the affinity of DNA sequence fragment and histone determines whether a DNA sequence fragment can wrap around a histone, they experiment in vivo and in vitro with DNA sequence.
fragment and histone of chicken [2]. Soon, this viewpoint extends by Field and Kaplan [3, 4].

So in this paper, we use both the affinity and flexibility to improve our research. First, we present our material and methods in this paper. Then, we show our experimental results and discussion. Finally, we conclude our research.

2. MATERIAL AND METHODS

Open data in this paper is from experimental result of N Kaplan etc. in [20], which is available in website (http://genie.weizmann.ac.il/pubs/nucleosomes08/nucleosomes08_data.html). In the data of synthetic oligonucleotides, we prefer synthetic oligonucleotides measured by microarray than measured by sequencing because microarray has higher accuracy. All data are created a pool of ~40,000 double-stranded oligonucleotides of length 150bp, and each combined with limiting amounts of chicken histones. Then the wrapped ones are extracted that had successfully competed to form nucleosomes. Finally, the affinity is calculated as the log-ratio between the reconstituted fraction and the initial pool as a measure. The results are calculated by oligonucleotides that were sequenced at least once and at most 500 times in each experiment.

Amount of Data is 43796 in this paper, which contains 25108 ‘positive’ sequence fragments (affinities of these sequences are positive) and 18688 ‘negative’ sequence fragments (affinities of these sequences are negative). We assume that the positive sequences are those who have high probability to combine the histone, and the negative ones have low probability to form nucleosome. Then, in order to winnow the data with lower properties, we choose the data with affinity more than 1 or less than 0. Then, we have 11539 positive sequences and 5221 negative sequences.

We first count the frequency of each diad in the data. The result are presented in Fig. (1), where left one is frequency of diad in all data, middle one is frequency in positive data and right one is frequency in negative data. We find that diad CC, CG, GC and GG occupy high frequency in core-DNA with positive affinity and AA, AT, TA, TT occupy high frequency in core-DNA with negative affinity. This is same to existing research result. We also have that diad AC shows high frequency in core-DNA with positive affinity.

Then, we calculate $w_i$ for each DNA sequence fragment $i$ in Eq.1, where $n_1i$ denotes the number of positive diad and $n_2i$ denotes the number of negative diad. Then, we divide all data to 10 parts equally and use neural network and leave-one-out method to process (only for study). In this case, we
get the weight result for every position. Moreover, we have three experimental results to validate our conclusions, where the only difference between them is the different chosen diad. We reach the mixed result in the following section.

\[ w_i = \begin{cases} \frac{af_i}{n_{1i} - n_{2i}}, & \text{if } a_i \geq 0 \\ \frac{af_i}{n_{2i} - n_{1i}}, & \text{if } a_i < 0 \end{cases} \]  

(1)

3. RESULTS AND DISCUSSION

First, we use the training data to reach the key position in these core-DNAs for three experiments. These three experiments are different from each other only by their diad. In first experiment, we use diad CC/CG/GC/GG as positive training data set and diad AA/AT/TA/TT as negative training data set. In second experiment, we drop diad TA in the negative data set because the frequency of diad TA does not show significant differences between positive and negative datasets. In third experiment, we add diad AC as a positive comparison because the diad AC also shows significant differences between positive and negative datasets. Then, the training results of these three experiments, which are the training weights of all positions, are shown in Fig. (2).

From Table 1, we find that the best positions of the 3 experiments are all similar. If we add the positions with largest weights to 20 in Table 2, we find that they are similar to each other with some same properties. So we believe that these properties may be suitable for other species. On the other hand, though our results are not so well than other results with yeast, we have to say that genome in chicken is more complex than in yeast. So we think our result is meaningful enough.

After that, we show our 3 experimental result in Table 1, where the only different between them is the chosen diad. To make the paper simplify, we rename them to 4-4, 4-3 and 5-4 instead. The Accuracy (Acc), Sensitiveness (Sen), positive predictive value (PPV) and Matthews’s correlation coefficient (MCC) denote the 4 most important indexes in the experiment. The formulas of these 4 indexes are presented in Eqs.2-5, where true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are all known in these experiments. In the first experiment, we have TP=786.3, FP=367.5, TN=1763.1 and FN=105.6 (mean of ten times with 9-1 model). In the second experiment, we have TP=852.6, FP=301.2, TN=1678.2 and FN=190.5. In the third experiment, we have TP=859.8, FP=294, TN=1673.7 and FN=195.0.

\[ \text{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \]  

(2)

\[ \text{Sen} = \frac{TP}{TP + FN} \]  

(3)

\[ \text{PPV} = \frac{TP}{TP + FP} \]  

(4)

\[ \text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}} \]  

(5)

Finally, we compare the weights of all positions in all these three experiments in Fig. (3). We also use only positive dataset and only negative dataset in the three experiments to
Table 2. 20 Key Positions in Experiments.

<table>
<thead>
<tr>
<th>4-4</th>
<th>1,2,22,33,42,43,44,53,54,64,65,74,75,85,86,95,106,117,127,128</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-3</td>
<td>2,13,14,37,45,46,54,65,66,75,84,86,91,105,106,113,117,118,121,132</td>
</tr>
<tr>
<td>5-4</td>
<td>1,2,11,12,22,43,46,53,54,55,56,66,75,76,85,86,87,96,117,127</td>
</tr>
</tbody>
</table>

Fig. (3). Comparison of the training weights in all positions for the three experiments, the blue line is for 4-4, red line in for 4-3 and green line is for 5-4.

Fig. (4). Comparison of the training weights in all positions for the three data sets in each experiment, the left sub-figure is for 4-4, the middle sub-figure is for 4-3, the right sub-figure is for 5-4, in each sub-figure, the blue line is for positive + negative data sets, red line in only for positive data set, 4-3 and green line is for negative data set.

compute the weights of positions in Fig. (4). From Figs. (3-4), we find that all these three experiments reach similar weights of positions, and the weights of positions are also similar between the three datasets we chosen. For example, in Fig. (3), we find that the tendency and extreme points are all similar between the three lines. However, in Fig. (4), we find that the maximum points in each sub-figure are similar between the blue and red lines, and the minimum points in each sub-figure are similar between the blue and green lines. So this means that the key positions are existed as a natural property of core-DNA.

CONCLUSION

In this paper, we presented a novel method to predict nucleosome positioning. In this method, we used a novel thinking with both key positions and affinity. We use Segal’s data to find the key positions and use these positions only to predict if a DNA sequence fragment is a core-DNA. Experimental results showed its effectiveness. Next step, we will divide each DNA sequence fragment to three parts since we know the combined positions in core-DNA is just the middle positions. So we will divide each DNA sequence fragment (150bp) to front (1-30), mid (31-120) and last (121-150). Then we will experiment if there will be better results.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous reviewers for their helpful comments in reviewing this paper.

This work is supported by grants Postgraduate Scientific Research Innovation Foundation of Inner Mongolia [B20141012610Z], Programs of Higher-level talents of Inner Mongolia University [125126, 115117], Scientific projects
of higher school of Inner Mongolia [NIZY13004], National Natural Science Foundation of Inner Mongolia [2014BS0606], National Natural Science Foundation of China [61261019, 61262082].

REFERENCES