One-base-pair Localization of Nucleosomes

Xiuwen Liu, Yu Zhang
Department of Computer Science
Florida State U., Tallahassee, FL, USA
{liux, yzhang}@cs.fsu.edu

Teng L. Liu
James S. Rickards High School
Tallahassee, FL, USA
tll10@my.fsu.edu

Jonathan H. Dennis
Department of Biological Science
Florida State U., Tallahassee, FL, USA
dennis@bio.fsu.edu

Abstract—Nucleosome position plays a key role in eukaryotic gene regulation, and its precise measurement is critical to understanding all DNA-templated events, but insufficient information has been available to infer nucleosome position at single-base-pair (1-bp) resolution. Brogaard et al. therefore developed an in vivo nucleosome-mapping method, based on the direct incorporation of a label into the nucleosome core, that generates hydroxyl radicals that cleave nucleosomal DNA at defined positions. This method allows direct measurement of nucleosome position with unmatched accuracy. Here we present a new algorithm that uses the Brogaard et al. data to achieve 1-bp resolution by using a joint cleavage pattern of radicals on Watson and Crick strands and can localize nucleosomes that are adjacent to each other. This algorithm provides the highest-resolution nucleosome-position maps to date and will allow detailed analysis of chromatin organization and protein-interaction dynamics.

Keywords—nucleosome; nucleosome localization;

I. INTRODUCTION

The importance of nucleosome position in the regulation of genomic processes and chromatin organization is well established (e.g., [3], [4]). In order to test and evaluate various hypotheses, nucleosome positions must be observed as accurately as possible. One commonly used procedure is to digest linker DNA using micrococcal nuclease (MNase). However, it is not feasible to control the digestion process so that nucleosomal DNA is preserved precisely. Even though short sequencing can potentially map digested nucleosome positions onto the genome accurately, the nucleosome position resolution is inherently limited. Although quantitative and statistical models have been proposed (e.g., [2]), dynamic aspects of nucleosome positioning and therefore occupancy could not be determined because nucleosome position could not be determined precisely enough.

Recently, Brogaard et al. have succeeded in generating data that contain genome-wide positioning information at 1-bp resolution by combining cleavage specificity of hydroxyl radicals and deep sequencing [1], but their proposed nucleosome-localization algorithms cannot achieve 1-bp resolution because of local smoothing. Here, we propose an original algorithm based on the crucial new observation that cleavage sites form two distinct groups and that any given nucleosome is strongly associated with one or the other - that can do so.

II. NUCLEOSOME LOCALIZATION ALGORITHM

A. Experimental Datasets

By incorporating mutant histone H4 onto nucleosomes, hydroxyl radicals cleave the DNA over two specific small regions relative to the nucleosome center, where the majority of the cleavage sites occur at two primary sites and two secondary sites. The process results in DNA segments, which are then sequenced to recover the cleavage sites; see [1] for more information about the datasets and experimental protocols.

The experiments in [1] generated six datasets, four are single-end sequenced and two are paired-end sequenced with varied incubation time. We have used the four sets of data by single-end sequencing. Note that a requirement to achieve single-base positions of nucleosomes is that the short sequences must be mapped onto the genome accurately as sequencing and mapping errors will result in errors in nucleosome position localization. We independently checked the sequencing errors of the four sets. There are in total of 4,500,569 reads on the Watson strand, 4,408,380 of which are perfectly mapped onto the genome (which is 97.95% of the total) and all the others have only a single mismatched nucleotide (which is 2.05%); similarly, there are in total of 4,480,929 reads on the Crick strand, 4,390,288 are perfectly mapped onto the genome (which is 97.98%) and all the others have also only a single mismatched nucleotide. There is no single short sequence with two or more mismatched nucleotides. These statistics indicate the high quality of sequencing of short sequences.

B. Localization Algorithm

As shown in [1], each nucleosome presents four possible cleavage sites: primary (PW) and secondary (SW) on the Watson strand and corresponding primary (PC) and secondary (SC) on the Crick strand. A nucleosome center can be localized if the cleavage sites can be classified correctly as these cleavage sites are known locations relative to the associated nucleosomes; using the 5' to 3' as the positive direction, a nucleosome is located at +1, -6, -1, and +6 nucleotides from PW, SW, PC, and SC respectively.

PW and SC are physically closer together and so are PC and SW, suggesting that PW and SC, and PC and SW may...
be strongly associated with each other as a pair. We therefore first identified all the local peaks on the Watson strand by finding all the reads that are larger than their immediate neighbors and are above 80th percentile but below 99.75th percentile of all the reads on the strand. We chose 80th percentile as the low threshold as it gives a good estimate of possible nucleosomes. We also ignored reads above 99.75th percentile because some locations have very high reads due to repeats at those locations. Additionally, as shown in [1], due to the size limit of the deep sequencing machine, bands no more than 200 nucleotides are isolated and used, causing an asymmetry at cleavage sites flanking to nucleosome free regions (which result in much longer fragments and therefore are not isolated). We also excluded the local peaks if there is no read in the 21 bp window on the Crick strand higher than the reads at 80th percentile on the Crick strand and separate average profiles are computed at these local peaks. Figure 1 shows the reads on Watson and Crick strand for a region on Chromosome 7 with identified local peaks; clearly these local patterns are consistent with the average local patterns, even though the variations are much larger.

Using the identified local peaks, we computed the average of the reads in a neighborhood of 21 to 21 bp around the center on the Watson strand and the corresponding average on the Crick strand. The average patterns at these sites are shown in Fig. 2(a) and the average patterns at sites that flank to nucleosome free regions are shown in Fig. 3(a). We also repeated the procedure for the Crick strand and the results are shown in Fig. 2(b) and Fig. 3(b) respectively. Because reads on the Watson strand and Crick strand are very symmetric overall, the identified patterns are indeed very close as expected. To verify that the local patterns do not depend on the choice of the percentile parameters, we repeated the procedures with low threshold percentile from 80th to 90th and high threshold percentile from 99th to 99,999th, the local patterns are consistent.

![Figure 1](image-url)

Figure 1. Reads on Watson and Crick strands from 319,270 to 319,340 on Chromosome VII, showing the read counts (top) and identified local peaks (bottom). Here the reads and identified local peaks on the Watson strand are plotted in blue and on the Crick stand in red.

These plots show clearly that a nucleosome is strongly associated with PW or PC but not both. For example, if a nucleosome is associated with PW, then the corresponding primary cleavage site on the Crick strand shows a peak even lower than the associated SC; similarly, if a nucleosome is associated with PC, then the corresponding primary cleavage site on the Watson strand shows a peak even lower than the associated SW. Clearly a nucleosome is strongly preferably associated with either PW or PC cleavage site but not both; therefore a nucleosome could be classified as PW-high (PWH) or PC-high (PCH) based on the dominated primary cleavage site. If there were no biases, there should be equal peaks at both primary sites.

On this basis, we have devised a family of inference algorithms to recover the precise positions of the nucleosomes and estimate the nucleosome occupancy score for identified nucleosomes. Our basic idea is to classify and estimate the reads associated with the candidate nucleosome with each basepair. Initially we classify all the candidates independent of other locations. For each location, we compute a likelihood score assuming that it is a PWH; the score is computed by comparing the reads at -1, +1, +6, and -6 with the identified local patterns shown in Fig. 2(a), relative to the nucleosome that is centered at 0 (with the 5' to 3' direction as the positive). One way is to compute the correlation between the local reads and the average; however, while correlation works well for a local pattern is close to the average pattern, correlation is not defined if all the reads are the same. To overcome this problem, we use the Euclidean distance between the patterns by normalizing the sum to 1, essentially treating them as probability distributions; any other distance or divergence between two probability distributions can be used. The probability distribution based distance however does not take into account the number of reads, where more reads indiate a stronger nucleosome. In general, we can pose the problem in the Bayesian inference framework. In this paper we have approximated the number of reads by a Heaviside step function. If the number of reads at a site is in the upper quantile, then it is considered a nucleosome center; otherwise, the reads are assumed due to noise. Note that used threshold can be changed easily. Similarly, we also compute the likelihood score assuming that is a PCH, where the score is computed by comparing the reads at -1, +1, +6, and -6 but with the identified local patterns shown in Fig. 2(b). A candidate is classified as PWH or PCH based on the computed likelihood score. Note that a candidate that does not have enough reads on both strands will have a likelihood score of zero.

With all the candidate nucleosomes classified, we need to resolve ambiguities and conflicts among the associated cleavage sites. The problem can be posed as an optimization one on a directed graph, where a nucleosome has two states (PWH and PCH) and directed connections to cleavage sites that support the assigned state, which implicitly classify all the associated cleavage sites. For example, if a nucleosome (centered at 0) is classified as PWH, then we have PWH at -1, SCH at -6, PCL at +1, and SWL at +6; on the other hand, if a nucleosome (centered at 0) is classified as PCH, then we have PCH at +1, SWH at +6, PWH at -1, and SCL at -6. Clearly some cleavage sites may have
conflicted classifications. By postulating that a cleavage site on the Watson strand can only be PWH, SWL, PWL, or SWH, similarly a cleavage site on the Crick strand can only be PCL, SCH, PCH, or SCL, in this paper we use a greedy algorithm by identifying the nucleosomes according to likelihood scores from the highest until a threshold is reached (0.3 is used in this paper so that the ones that flank to nucleosome free regions will be identified as well without a separate step). We first identify the one with the highest likelihood score and we iteratively identify the one with the highest likelihood score among the remaining candidates under the constraint each cleavage site can be classified as one of the four classes. On each strand, the first two on each strand are associated with PWH and the last two with PCH. Other optimization algorithms can also be used. Note that nucleosome free regions will divide each chromosome into multiple connected components and the optimization needs to be done on each connected component. The effectiveness of such optimization algorithms is being investigated.

![Figure 2](image1.png)

(a) Average cleavage patterns around local peaks, the blue curve shows the average reads on the Watson strand and the red curve on the Crick strand; local peaks are identified on the Watson strand (a) and on the Crick strand (b).

![Figure 3](image2.png)

(a) Average joint cleavage patterns around local peaks that flank to nucleosome free regions, the blue curve shows the average reads on the Watson strand and the red curve on the Crick strand; local peaks are identified on the Watson strand (a) and on the Crick strand (b).

Note that after nucleosomes are identified, we can obtain a more accurate estimation of the local patterns by averaging over PWH nucleosomes and separately PCH nucleosomes. As these sites are more specific than the local peaks used in Figs. 2 and 3, the estimated patterns should be more pronounced. The updated patterns can be used to reestimate the nucleosomes; it seems that the identified nucleosomes do not change significantly as the initial local patterns are sufficiently accurate.

III. EXPERIMENTAL RESULTS

We have applied the greedy algorithm to the combined reads on the 16 yeast chromosomes. Figure 4(a) shows that the algorithm could identify all the nucleosomes in the region associated with significant local peaks. Figure 4(b) illustrates the key difference between the proposed algorithm and the weighted averaging algorithm proposed in [1]. Even though the two nucleosomes are immediately adjacent, our algorithm correctly identifies both, whereas the algorithm in [1] identifies the average of the two.

![Figure 4](image3.png)

(a) (b) Figure 4. Two examples of nucleosomes identified by the proposed example, where the identified nucleosomes are shown in colored vertical bars. In (b), for clarity, identified nucleosomes by the proposed algorithm are given in the middle panel and the one by Brogaard et al. [1] at the bottom panel.

IV. CONCLUSION

The ability to localize nucleosome centers to 1-bp resolution by means of our new algorithm enables us to characterize new dynamic nucleosome properties, such as motion flow of nucleosomes, which may lead to new insights to chromatin organization as nucleosomes are the essential building blocks. In addition, by posing the problem as one of global optimization, we can compute multiple solutions and also allow multiple classifications of a cleavage site, which are being investigated.

Acknowledgment The authors would like to thank Dr. Ji-Ping Wang for his help in understanding the experiments in [1].

REFERENCES


