A dynamic programming algorithm for nucleosome positions alignment

Yiru Zhang, Changchang Cao, Hongde Liu, Xiao Sun*
State Key Laboratory of Bioelectronics
School of Biological Science & Medical Engineering, Southeast University
Nanjing, China
E-mail: zhangyiru_seu@163.com

***Abstract*—Nucleosomes are the basic units of eukaryotic chromatin. The nucleosome positioning is dynamic for various cell types and biological states, resulting in specific gene regulation. Currently, there is no approach to find the correspondence between two sets of nucleosomes to reveal the difference of their positions.**

We develop a method for nucleosome positions alignment based on the dynamic programming algorithm, which can quantify the changes in nucleosome locations with scores and evaluate regional dynamics changes including translation and missing. Given the result of a peak list stands for nucleosome positions, to align the peaks from two samples, our method accumulate all pair scores for match, replacement or deletion and choose the maximum one as the optimal alignment. From nucleosome alignment we can find one-by-one correspondence between nucleosome positions in different cell stages and the conservative stable and variable regions, which can be used to recognize dynamic behaviors of nucleosome shift and eviction.

Keywords-nucleosome positioning, dynamic programming algorithm, transcription regulation.

I. INTRODUCTION

In eukaryotic cells, the nucleosome is the basic unit of chromatin, which wraps 147-bp DNA around an octamer of histones. The nucleosome organizations have been implicated to make nucleotide a density varies rosary shape to form chromatin structure in regulating DNA replication, DNA repair and gene transcription [1, 2]. In recent years, it has become a hot spot to study the functions of the changes in chromatin structure and nucleosome positioning in transcriptional regulation [3, 4, 5].

Dynamic nucleosome positioning can pack or release DNA fragment for proteins contact. Compared with the DNA binding sites packed in nucleosomes, it is easier for transcription factors to recognize them outside. When the cell state changes, nucleosomes may relocate to remodel the chromosome structure and expose different transcription factor binding sites, TFBSs. Thus, transcription factors will bind DNA and regulate the gene-specific expression. In addition, nucleosome free region (NFR) in the upstream of transcription start site (TSS) is conducive for the transcription initiation complex combination to regulate gene expression [6]. From another perspective, the collapse and reshape of nucleosomes can make the chromatin topology spatial conformation changes for transcription factor binding and RNA polymerase II elongation.

Chromatin immune precipitation-chip (ChIP-Chip) and chromatin immune precipitation-sequencing (Chip-Seq) [5] are two important methods to infer the nucleosome positions. Recently, a new technique called micrococcal nuclease-sequencing (MNase-Seq) [7] was invented which could also be used to obtain the nucleosome positions. Following by co-immunoprecipitation and DNA sequencing, all these methods will get short sequence reads which can be mapped back to reference genome to find nucleosome positions.

It has been gradually recognized that nucleosome positioning is dynamic for different cell types or biological states. Previous studies mainly focused on specific functional regions, such as promoters, for studying the relationship of dynamic nucleosome positioning and transcriptional regulation. For instance, Field et al. found nucleosome deletion in the upstream of transcription start site by analyzing the yeast profiles [9, 10]. Zaugg and Luscombe studied and proposed a qualitative model for nucleosome architecture in yeast promoters. The model consists of two NFR configurations, Open and Closed, and two expression states ON and OFF. Their model demonstrated that the promoter nucleosome organization influences the expression state, not the expression level [6].

Recently, two methods were proposed to compare the different nucleosome positioning from two samples. One is DiNuP which was developed by Zhang et al. http://www.tongji.edu.cn/~zhanglab/DiNuP [13]. DiNuP combined reads from two samples with sliding window. The Kolmogorov-Smirnov (K-S) test was then used to calculate the difference between these re-sampled windows. A statistical P-value was provided for each identified region of differential nucleosome positioning (RDNPs) based on the difference of read distributions. DiNuP was shown to be both sensitive and specific for the detection of changes in nucleosome location such as occupancy and fuzziness. Chen et al. provided another method called DANPOS based on Poisson distribution for two samples [25], http://code.google.com/p/danpos/. Sample with higher nucleosome occupancy was used as preference (Poisson distribution parameter λ), while the other sample was used to calculate the difference. The result was classified into three categories including nucleosome position shifts, fuzziness changes, and occupancy changes.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

However, both DiNuP and DANPOS ignored the biological significances of different kinds of nucleosome dynamic changes which may results in incorrect alignment. In reality, the nucleosome dynamic shift or eviction will result in different biological processes.

In this paper, we designed an algorithm for nucleosome positions alignment based on dynamic programming. Our method could quantify the nucleosome reposition with designed scores and evaluate regional dynamics changes including replacement and deletion. Several methods have been proposed to identify whole genome nucleosome positions by analyzing Chip-seq reads distribution [22, 23, 26], which will produce a peak list standing for the nucleosome positions. After all pair scores for match, replacement or deletion are accumulated, we can find one-by-one correspondence between nucleosome positions in different cells. From which we can identify dynamic behaviors of nucleosomes such as shift and eviction, and recognize conservative stable and variable regions.

II. MATERIALS AND METHODS

A. Data

The coordinates for all uniquely mapped sequencing reads of nucleosome in resting and activated human CD4+ T cells is downloaded from NIH website

http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx.

B. Identify nucleosome positions

Several methods have been proposed to identify whole genome nucleosome positions by analyzing Chip-seq reads distribution [22, 23, 26], and we chose MACS to get the nucleosome occupation signal [26], which first map the short reads back to the reference genome, then extend them to 147bps for overlapping and modeling Poisson distribution (Fig.1.A).

To find individual peak positions in overlapped signals, maximum spectrum of continuous wavelet transform (MSCWT) was utilized to identify the center of each peak as the nucleosome location [27] in our research (Fig.1.B).

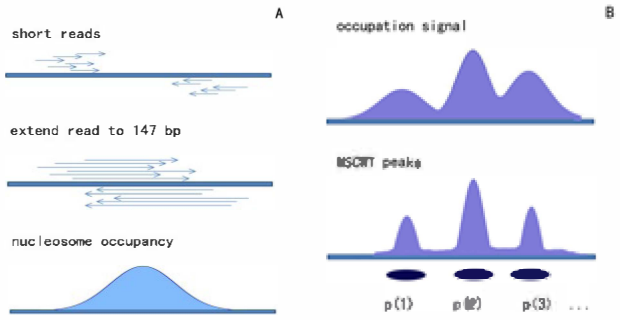

Fig.1. Framework for identifying nucleosome positions. (A) The workflow of MACS, which is used to calculate the nucleosome occupancy. (B) Maximum spectrum of continuous wavelet transform (MSCWT) is used to recognized individual peaks in overlapped signals.

C. Nucleosome Positions Alignment

In nucleosome occupation signal profile, each peak indicates a nucleosome. We use a peak list to represent nucleosome positions. Peak lists P and Q consist of peaks which stand for nucleosome positions from two different samples:

(1)

Peak list P contains m nucleosomes and Q has n. The alignment of two peak lists refers to a one-by-one correspondence which indicates the changes between these two lists of nucleosomes. Nucleosomes with the same location will be considered as match, while nucleosomes which is missing or shifting will be considered as dynamic positioning.

For instance (Fig.2.A), there are three pairs of nucleosomes that are matched, p(1)-q(1), p(2)-q(2), p(3)-q(3). However, both nucleosome shift and missing could occur when dealing with p(4) and q(4). To solve this problem, dynamic programming is applied to calculate the better correspondence by transforming the peak sequence alignment as a series of single peak sub-alignments, from the first nucleosome to the last one. This step will be repeated to find the optimal alignment.

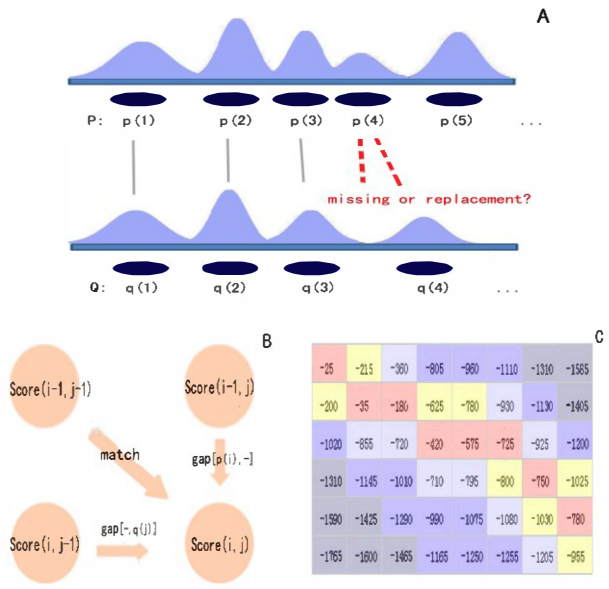

Fig.2. Find the optimal alignment between peak lists P and Q. (A) With the same genome, the nucleosome positions from two samples may have corresponding relationship. With different biology significance, replacement or missing should be scored differently. (B)The object is to calculate the best score for each position in a matrix with sequences running across the first to the last. (C) Eventually, all possible combinations of sequence matches and gaps are taken into account. The highest-scoring red position gives the corresponding sequence alignment [28].

Consider prefix P_i and Q_j, which contains all peaks in the front of p(i) and q(j) respectively. Score(i, j) is the highest score for the two prefix alignment , and there are three possible correspondence between p(i) and q(j) as follows:

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

- Starting from Score(i-1, j-1) for the prefix Pi-1 and Qj-1 , to match p(i) and q(j).
- Starting from Score(i-1, j) for the prefix Pi-1 and Qj , to align p(i) with a gap where a nucleosome is missed in Q.
- Starting from Score(i, j-1) for the prefix Pi and Qj-1 , to align q(j) with a gap where a nucleosome is missed in P.

We score for each correspondence and choose the maximum one as Score(i, j) .

(2)

Match [p(i),q(j)] denotes the score for the match of p(i) and q(j). We score it according to the distance between the coordinates of p(i) and q(j) . The nearer they are, the higher the score will be.

(3)

A gap means at least one nucleosome is missed in one of these two samples. For example, Gap [p(i),-] indicates that a nucleosome is missed in sequence Q at the corresponding location of p(i). The score of Gap is formulated as below. After the alignment of prefix $P_{i-1}$and Q_j, if there is no nucleosome in sequence Q from the location of p(i-1) to p(i), we consider a gap exists and score it with its size.

(4)

However, for a long segment of DNA without wrapping histones to form nucleosome in the genome which may results in a large gap with a very low score, the algorithm will match the next nucleosome immediately to make sure the score is not too low, which will produce imperfect results(Fig.3.A).

To fix this bug, we set -147 as the maximum penalty for Gap, which equals to the length of the DNA in a nucleosome. When the space between two samples is more than this value, the penalty remains as -147. (Fig.3.B).

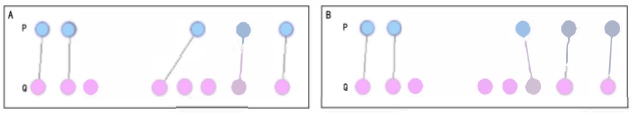

Fig.3. Larger gap will cause matching errors. (A)With a large gap in front of p(3) and q(4), it will choose match score immediately in order to avoid the score being much lower. Obviously, it is not the optimal alignment. (B)The result is better when set147 as a maximum penalty which means no matter how large the gap is.

In summary, the gap score is formulated as below:

(5)

Starting at the left-upper, i=0, j=0, to the right-downer, i=m, j=n, eventually, all possible combinations of sequence matches or gaps are taken into account. Score (m, n) is for the global alignment, from which we can trace back to find the optimal nucleosome alignment with the highest score (the red pathway in Fig.2.C). From the results, we could get the best correspondence between nucleosome positions from two cells and infer the conservative stable or different variable regions easily.

III. RESULT AND DISCUSSION

We used the nucleosome positions data of chromosome 20 from CD4+ activated and resting cells to conduct simulation experiment (Fig.4). The abscissa indicates the coordinate of genomic DNA sequence. Blue circles represent the nucleosomes in activated cells, while purple for the resting cells. Straight lines connect the matched pairs of nucleosomes, and the black curve below shows the degree of differences between nucleosome positions from these two cells.

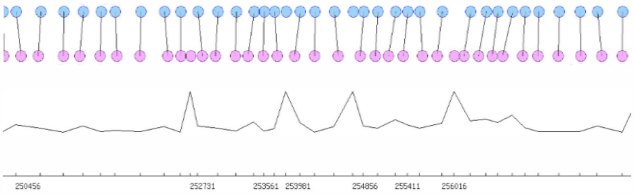

Fig.4. The alignment result of nucleosome positions from activated and resting CD4+ cells. The abscissa indicates the coordinate of genomic DNA sequence. Blue circles represent the nucleosome positions in activated cells, while purple for the resting cells. Straight lines connect the match pairs. The black curve shows the degree of differences between two samples. The curve is smooth if the nucleosome location shifts little, which is conservatively positioning. On the contrary, rising curve reflects larger distance and a peak may indicate one nucleosome missing.

TABLE I. SOME GENE EXPRESSION DATA IN ACTIVATED AND RESTING CD4 + CELLS

chr	Gene	TSS	TTS	ID_mRNA	State	Value ABS_call	P-value
20	VAPB	56397651	56455369	NM_004738	Resting	3055.42	0.000732422
					Activated	3222.28	0.000244141
20	ADNP	48980934	48940290	NM_015339	Resting	5739.01	0.000244141
					Activated	5448.21	0.000244141
20	SEC23B	18436188	18490050	NM_006363	Resting	1853.37	0.000244141
					Activated	7131.19	0.000244141
20	ENTPD6	25124372	25155360	NM_001247	Resting	2333.89	0.00195313
					Activated	1039.64	0.00195313

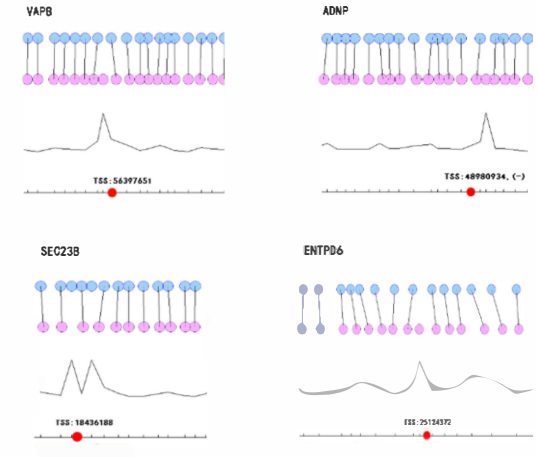

Fig.5. The alignment result of some genes, representatively, VAPB、ANDP、SEC32B、ENTPD6. There is one or two missing gap in the upstream of transcription start site, TSS, while the downstream has perfect corresponding.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

From the calculation result, we can infer the site of reposition easily, especially the missing nucleosomes. Combined with the gene expression data of activated/resting CD4+ cells, the results showed that only one or two missing nucleosome in the upstream of transcription start site, while the downstream had perfect corresponding (table 1, Fig. 5). This results is consistent with previous researches: nucleosome reposition or deletion happens in promoter region to regulate the gene-specific expression when the cell state changes.

IV. CONCLUSION

In this paper, taking advantage of dynamic programming, we proposed a method for nucleosome positions alignment, which accumulated all pair scores for match, shift and missing. By computing the high-score alignment path, we could obtain the optimal correspondence between nucleosome positions of two samples. From the results, we can infer the differences and find the dynamic regions between different samples. Given the function of genomic region, the biological significance could be further analyzed.

Combined with the distribution of functional DNA elements, further efforts will be paid in studying the behavior mechanisms of reposition or deletion in the future.

ACKNOWLEDGMENT

This work was supported by the National Basic Research Pr ogram of China (2012CB316501) and the National Natural Sci ence Foundation of China (61073141).

REFERENCES

[1] Bruce Alberts/Alexander Johnson/Julian Lewis/Martin Raff/Keith Roberts/Peter Walter. [M] MOLECULAR BIOLOGY OF THE CELL, 5th Edition，50~20.

[2] Kaplan N, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009，458 (7236):362-366.

[3] Field Y, et al. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol. 2008，4 (11):e1000216.

[4] Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature. 2011, 474(7352):516-20.

[5] Yi X, Cai Y-D, He Z, Cui W, Kong X，Prediction of Nucleosome Positioning Based on Transcription Factor Binding Sites. PLoS ONE, Vol. 5 Issue 9, p1-7. 7p.

[6] Judith B. Zaugg and Nicholas M. Luscombe, A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast, Genome research, 2012 22: 84-94.

[7] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K._Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008, 132(5):887-98.

[8] Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JZ, Widom J, A genomic code for nucleosome positioning Nature 2006 ,442 :772-778.

[9] Schwabish,M.A. and Struhl,K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. Mol. Cell Biol.2004，24, 10111–10117.

[10] Valouev,A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res., 2008, 18, 1051–1063.

[11] Bai L, Morozov AV. Gene regulation by nucleosome positioning. Trends Genet. 2010 Nov; 26(11):476-83.

[12] Sadeh R, Allis CD. Genome-wide "re"-modeling of nucleosome positions. Cell. 2011, 147(2):263-6.

[13] Fu K, Tang Q, Feng J, Liu XS, Zhang Y, DiNuP: DiNuP: a systematic approach to identify regions of differential nucleosome positioning. Bioinformatics, 2012, 28(15):1965-1971.

[14] Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009, 10(3):161-72.

[15] Polishko A, Ponts N, Le Roch K G, et al. NORMAL: accurate nucleosome positioning using a modified Gaussian mixture model [J]. Bioinformatics, 2012, 28(12): i242-i249.

[16] Brogaard K, Xi L, Wang J P, et al. A map of nucleosome positions in yeast at base-pair resolution [J]. Nature, 2012, 486(7404): 496-501.

[17] Kuangyu Yen·Vinesh Vinayachandran · Kiran Batta·R.,Thomas Koerber,B.Franklin Pugh, Genome-wide Nucleosome Specificity and Directionality of Chromatin Remodelers, Cell,2012, 149(7):1461-1473.

[18] Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. Nature Genetics, 2009, 41:438~445.

[19] Zaugg JB, Luscombe NM. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. Genome Res. 2012, 22(1):84-94.

[20] Mark D Robinson, David P De Souza, Woon Wai Keen,Eleanor C Saunders, Malcolm J McConville, Terence P Speed,Vladimir A Likić. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. BMC Bioinformatics 2007, 8:419.

[21] Yong Zhang, Hyunjin Shin, Jun S Song, Ying Lei and X Shirley Liu, Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq，BMC Genomics, 2008, 9:537.

[22] Christiana Spyrou, Rory Stark, Andy G Lynchand Simon Tavaré，BayesPeak: Bayesian analysis of ChIP-seq data，BMC Bioinformatics，2009, 10:299.

[23] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers & Wing H Wong, An integrated software system for analyzing ChIP-chip and ChIP-seq data, Nature biotechology, 2008 Nov;26(11):1293-1300.

[24] Hongde Liu, Xueye Duan, Shuangxin Yu, Xiao Sun，Analysis of nucleosome positioning determined by DNA helix curvature in the human genome，BMC Genomics 2011, 12:72.

[25] KaiFu Chen, Yuanxin Xi, Xuewen Pan, DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing,Genome Res. 2013 23: 341-351.

[26] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute，Model-based Analysis of ChIP-Seq (MACS)，Genome Biology 2008, 9:R137.

[27] Lu Xiaoquan, Liu Hongde, Xue Zhonghua, and Zhang Qiang，Maximum Spectrum of Continuous Wavelet Transform and Its Application in Resolving an Overlapped Signal，J. Chem. Inf. Comput. Sci. 2004, 44, 1228-1237.

[28] David W. Mount, Bioinformatics:Sequence and gomome Analysis,2006,56-57

978-1-4799-7294-4/14/$31.00 ©2014 IEEE