AN ASSESSMENT OF PREDICTION ALGORITHMS FOR NUCLEOSOME POSITIONING

YOSHIAKI TANAKA[1,2]
ytanaka@hgc.jp

KENTA NAKAI[1,2,3]
knakai@ims.u-tokyo.ac.jp

[1] *Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*
[2] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*
[3] *Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency, 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan*

Nucleosome configuration in eukaryotic genomes is an important clue to clarify the mechanisms of regulation for various nuclear events. In the past few years, numerous computational tools have been developed for the prediction of nucleosome positioning, but there is no third-party benchmark about their performance. Here we present a performance evaluation using genome-scale *in vivo* nucleosome maps of two vertebrates and three invertebrates. In our measurement, two recently updated versions of Segal's model and Gupta's SVM with the RBF kernel, which was not implemented originally, showed higher prediction accuracy although their performances differ significantly in the prediction of medaka fish and candida yeast. The cross-species prediction results using Gupta's SVM also suggested rather specific characters of nucleosomal DNAs in medaka and budding yeast. With the analyses for over- and under-representation of DNA oligomers, we found both general and species-specific motifs in nucleosomal and linker DNAs. The oligomers commonly enriched in all five eukaryotes were only CA/TG and AC/GT. Thus, to achieve relatively high performance for a species, it is desirable to prepare the training data from the same species.

Keywords: nucleosome positioning, prediction, sequence dependency, assessment

1. Introduction

Chromosomes of eukaryotes are composed of nucleosome arrays. By affecting the access of regulatory factors to DNA, nucleosomes are involved in various nuclear functions, such as transcription, replication and DNA repair.

It is known that nucleosome formation depends on local DNA sequence more or less. Several motifs favoring and inhibiting nucleosome formation have been detected by *in vivo* or *in vitro* experiments. As a typical instance, short nucleotides, such as AA/TT and GC, appear at about 10-bp interval in nucleosomal DNAs of several species [1, 2]. They are associated with the positions of the minor and major grooves facing with the histone surface and are associated with the bendability of DNA during nucleosome formation. On the other hand, several consecutive oligonucleotides, such as $(A)_n/(T)_n$ and $(CG)_n$, are reported to prevent DNA sequence from forming nucleosomes via their conformational change [2, 3]. It has also been reported that $(CG)_n$ promotes the formation of the left-handed Z-DNA conformation [4].

These sequence dependencies suggest the possibility of predicting nucleosome locations computationally. Recently, several prediction tools have been developed for the genome sequence of budding yeast and few other organisms [5-12]. For example, Segal *et al.* made position weight matrices that characterize periodic patterns of specific dinucleotides from about two hundred nucleosomal DNA sequences and scanned them with the dynamic programming method throughout the yeast genome [1]. Using this model, they suggested that about 50% of nucleosome placements are prefigured by genome sequence. Furthermore, they have recently updated their methodology by adding the preference of 5-bp short sequences in nucleosome-poor regions and have showed high correlations with *in vivo* and *in vitro* nucleosome distributions [11, 12].

Mielle *et al.* focused on the intrinsic bendability of DNA sequence and constructed a physical model of the DNA bending around the histone octamer [6]. Their method calculates the free energy of a DNA fragment to form the ideal curved structure without any training procedure. In promoter regions, the energy was highly correlated with the nucleosome occupancy, and the energy profile was conserved from *Drosophila melanogaster* to *Saccharomyces cerevisiae*.

Peckham *et al.* and Gupta *et al.* introduced Support Vector Machines (SVMs) to classify the nucleosomal and non-nucleosomal DNAs [10]. Using the 1,000 highest and the lowest scoring probes in nucleosome tiling arrays, they calculated the frequency of 2,772 all possible oligomers (from 1- to 6-mers). The SVM trained by these frequencies showed a good agreement with experimental data and they proposed several oligomers for the distinction between nucleosome formation and exclusion.

Thus, numerous tools for nucleosome positioning are currently available. However, users do not know which algorithm is more accurate in predicting *in vivo* nucleosome distribution or to which organisms these methods are applicable or inapplicable. To clarify these points, we evaluated the prediction accuracy of representative algorithms from three typical classes of prediction methods: Segal's methods [1, 11, 12], which are mainly based on the 10-bp sequence periodicity; Miele's method [6], which is based on the physical property of DNA; and Gupta's SVM method [10], which relies on the statistic of oligomer frequency. Using the genome-scale *in vivo* nucleosome maps in human, medaka fish, nematode, candida yeast and budding yeast, we show the difference of prediction performance among them.

2. Materials and Methods

2.1. Genome-Scale Nucleosome Maps

The genome sequences of four organisms (hg18, oryLat2, ce6 and sacCer1) were obtained from the UCSC database. The candida genome sequence was obtained from the website of Eran Segal's laboratory.

Raw nucleosomal DNA sequences in human (SRA000234), medaka (SRA002449) and nematode (SRA001023) were downloaded from the Short Read Archive database [13-15]. The human nucleosome tags in activated T cells were mapped to the genome by SeqMap [16], and the tags in medaka and nematode were mapped by MAQ with the allowance of 3-bp mismatch. Nucleosome tags mapped to the candida and budding yeast genomes were obtained from E. Segal's website [17]. Then, the uniquely mapped tags were used for determining the nucleosome locations by our own Hidden Markov Model (HMM), which uses the gradient of signals instead of intensity (Tanaka, Yoshimura and Nakai, submitted). As shown in the manuscript, the numbers and positions of allocated nucleosomes by our HMM were close to those of originally reported assignments. For parameter estimation of the HMM, 100 randomly selected regions of 10,000 bp, where at least one tag was observed at the 1000-bp interval, were used in each organism. Finally, we assigned 13,737,718, 3,480,027, 453,011, 54,971 and 56,172 nucleosomes in human, medaka, nematode, candida and yeast, respectively.

As an original training dataset in Gupta's method, we downloaded probe sequences in tiling array data provided by Ozsolak *et al*. from the website of William Stafford Noble's laboratory [18].

2.2. Application of Prediction Algorithms

As a positive and negative datasets, we randomly extracted 100 nucleosomal and 100 linker DNA sequences from the genome-scale nucleosome maps, respectively. In this study, we used only sequences whose lengths were from 100 bp to 200 bp. Since the nucleosome maps were constructed from short DNA tags, we did not use sequences containing repetitive elements. These processes were repeated 10 times.

We applied each prediction algorithm to 10 evaluation datasets. In Segal's method, we tested all of its three versions (ver. 1-3) with default parameters. Since the original authors recommend for users to add flanking regions for avoiding boundary effects, we added 5000-bp regions at both sides to the test datasets. If the scores within +/-10 bp of a nucleosome start site exceeded a cutoff value, we regarded that the nucleosome is correctly predicted by this model. For the evaluation of the ver.1, we used its yeast model.

In Miele's method, the unsigned average of all output free energies for each input fragment was used. If this value exceeds a cutoff score, we defined that the nucleosome is correctly predicted.

In Gupta's method, the discriminant value of SVM was used. We performed a 10-fold cross validation test: for the prediction of each test dataset, the other nine datasets were used for training parameters of SVM. Although a linear kernel ($p = 1$ and $\beta = 0$) was used in the original article, we tested additional five other kernels: quadratic ($p = 2$ and $\beta = 1$), cubic ($p = 3$ and $\beta = 1$) and three types of Radial Basis Function (RBF) kernel ($\gamma = 1, 5$ *and* 10), which we call RBF1, RBF5 and RBF10, respectively.

$$\text{Polynomialkernel}: K(x,z) = (\beta + < x, z >)^p$$

$$RBF : K(x,z) = \exp(-\|x - z\|^2 / \gamma)$$

2.3. Receiver Operating Characteristic (ROC) Curve

The fitness of each prediction method to experimental data was measured by the ROC curve and the Area Under the Curve (AUC) value. This measurement is useful for comparing the prediction performance of two or more tests simultaneously and graphically. The 0.5 of AUC is equivalent to random prediction. For each test, the curve was drawn by the averaged sensitivity/specificity of 10 evaluation datasets under various cutoff values, and the average AUC was used as a measure of the prediction accuracy.

2.4. Over- and Under-Represented Oligomers

Frequencies of 2,772-oligomers were calculated from 10 evaluation datasets as in Gupta *et al.* [10]. For each oligomer, its over- and under-representation in nucleosomal DNA compared with non-nucleosomal DNA were evaluated by one-tailed Welch *t* test.

3 Results and Discussion

3.1 *Prediction Ability of Each Algorithm for Overall Nucleosomes*

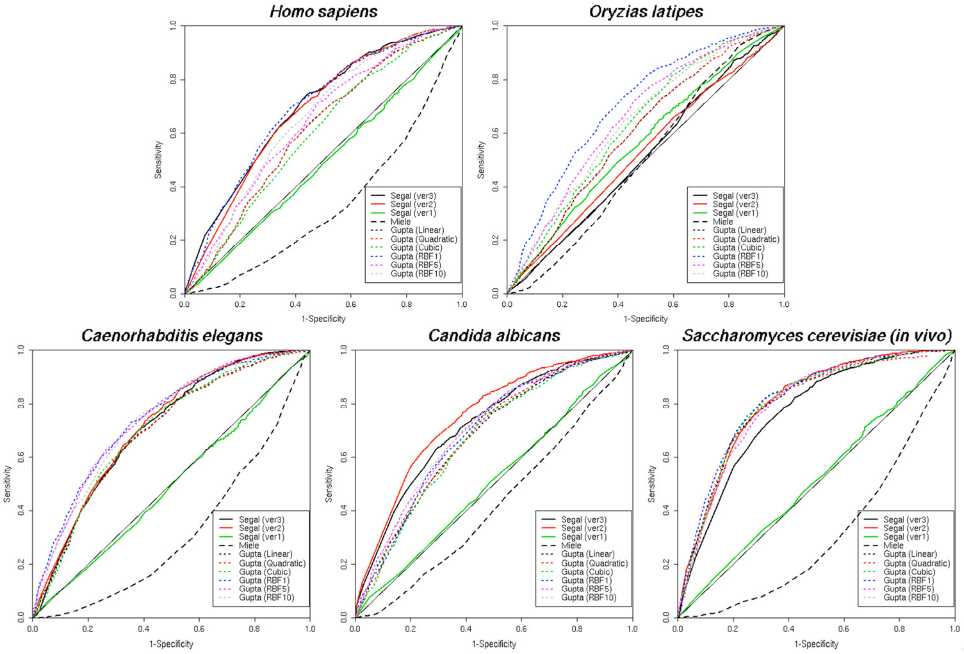

Figure 1: ROC curves of prediction methods for five organisms

Figure 1 shows the ROC curves of all three algorithms with several options. The results are rather varied for different organisms. In predicting nucleosomes of the three invertebrates, vers.2 and 3 of Segal's method show more than 0.700 AUC values (Supplementary Table 1). In particular, the highest AUC was observed in yeast. On the other hand, the AUC for medaka data was the worst, which was not largely different from random prediction. Except for medaka, the first version was worse than the other new versions. The two recently updated versions showed almost equal prediction accuracy. On average, the AUC of ver. 2 is slightly higher than that of ver. 3.

Miele's model did not work well in all organisms by our evaluation test.

Among the three algorithms, Gupta's SVM was the best in human, medaka, nematode and yeast. The RBF kernels were better than the polynomial kernels in all parameters and species. While the prediction accuracy for the three parameters was similar in each organism, RBF1 showed the highest AUC in human, medaka, and yeast nucleosomes in all six kernels. It was also the best on the average of all species. Comparing the prediction performance between Gupta's SVM with the RBF1 kernel and the ver. 2 of Segal's method, there were no significant difference in human, nematode and yeast. However, Gupta's SVM was significantly better in medaka while Segal's method was better in candida (*P < 0.001* by Wilcoxon test).

It is possible that the high prediction accuracy of Gupta's method is specific to the data source for model training. Therefore, we further analyzed the prediction performance of Gupta's SVM (RBF1) trained by Ozsolak *et al.*'s human tilling array data [18]. By applying the SVM trained by the Ozsolak data to the above evaluation dataset in human (Schones *et al.*'s data), the average AUC was 0.697, which was not significantly different from that of Gupta's model trained and tested by the Schones data (*P = 0.912* by Wilcoxon test; figure 2). This result indicates that the effect of specific training data to the performance of Gupta's method is not so significant.

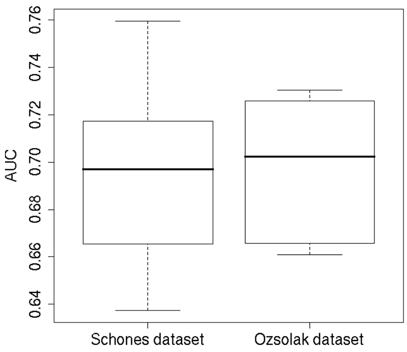

Figure 2. Comparison of prediction ability for Gupta's SVM by two different data

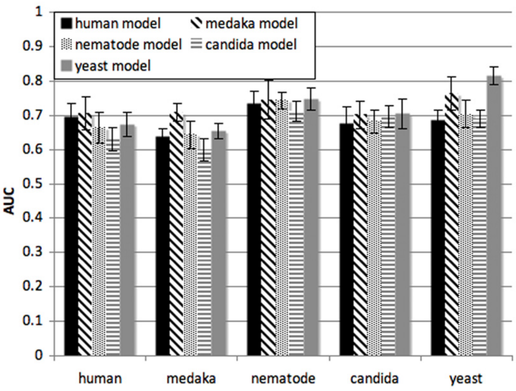

Figure 3. Prediction accuracies of Gupta's SVM trained by the data of various species

Additionally, we studied whether Gupta's model trained by a dataset of a certain organism was applicable to the prediction in other species. This test has two meanings: one is to assess its practical value when there is no training data available for the species that users want to analyze, while the other is to see if the sequence features used in the method are conserved across species. The SVM trained by the yeast dataset showed significantly greater performance in predicting itself than those trained by datasets from other species (*P < 4.19e-03* by Dunnett's test; figure 3). The model trained by the medaka data was significantly better in predicting themselves, too (*P < 3.23e-02*). On the other hand, there were no significant differences among the five training datasets for the prediction of nematode and candida nucleosomes (*P > 0.05*). These results suggest that the sequence specificity for nucleosome positioning may be a little varied in medaka and budding yeast.

3.2 *General and Specific Sequence Dependencies in Nucleosome Positioning*

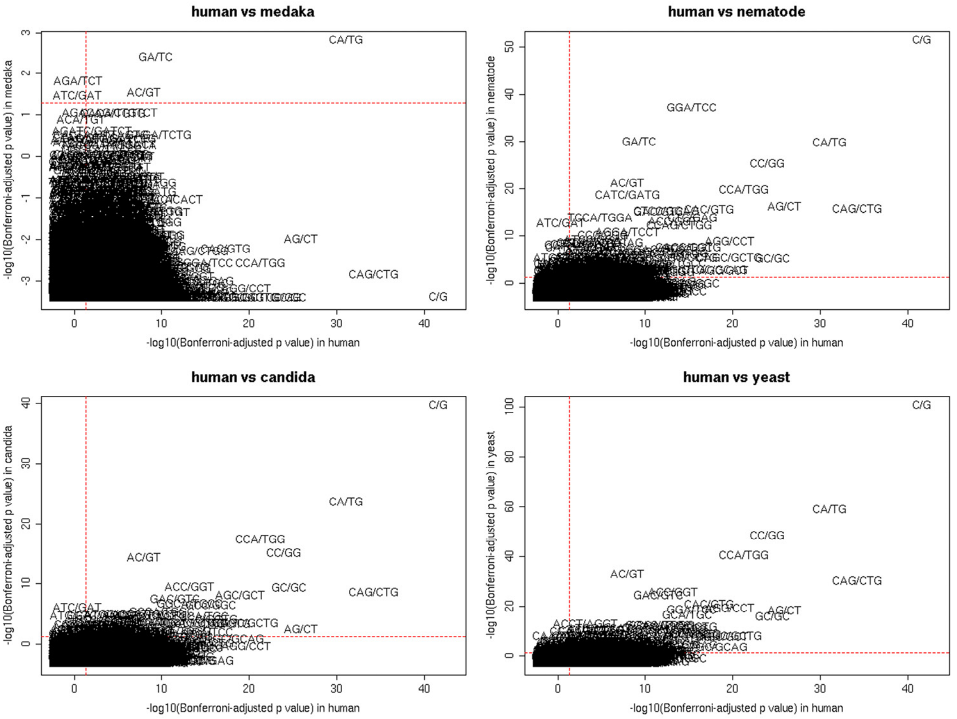

Figure 4: Comparative analyses of over-represented oligomers in nucleosomes. Significance level was set to 0.05 (indicated by dashed lines).

To further analyze the potential difference of DNA sequence dependency between species, we compared oligomers that are over-represented in nucleosomal DNAs between various pairs of species (Figure 4 and Supplementary Table 2). Clearly, the nucleosomal DNAs in medaka are quite different from others. Among the five over-represented motifs in medaka, CA/TG and AC/GT are shared in all species (Bonferroni-adjusted *P < 0.05*). The CA/TG step is known as the most flexible dinucleotide step [19], which is suitable for kinks of DNA to wind around a histone octamer [20]. AC/GT was also reported as a flexible step [19]. On the other hand, AGA/TCT is specific to only medaka, and there is no report about its association with nucleosome formation.

Except for medaka, a couple of additional general motifs were observed: C/G mononucleotide showed the lowest *p*-value in all tested oligomers. The over-representation of single nucleotides is interpreted that the G+C content of DNA sequence influence the nucleosome positioning. Furthermore, WW dinucleotides (CC/GG, GC and CG) were also over-represented. In addition, several trinucleotides were also observed in the four species: ACC/GGT, AGC/GCT, CAC/GTG, CAG/CTG, GAC/GTC, GCA/TGC and GGA/TCC, which are the combinations of one A/T and two G/C steps, were observed. The repetitive occurrence of CAG/CTG is known to form a stable nucleosome structure [3]. Gupta *et al.* also pointed out the 3-bp interval of CG and GC in human nucleosomal DNA [10]. It seems to be interpreted that proper spacing of A/T and G/C is important for promoting nucleosome positioning.

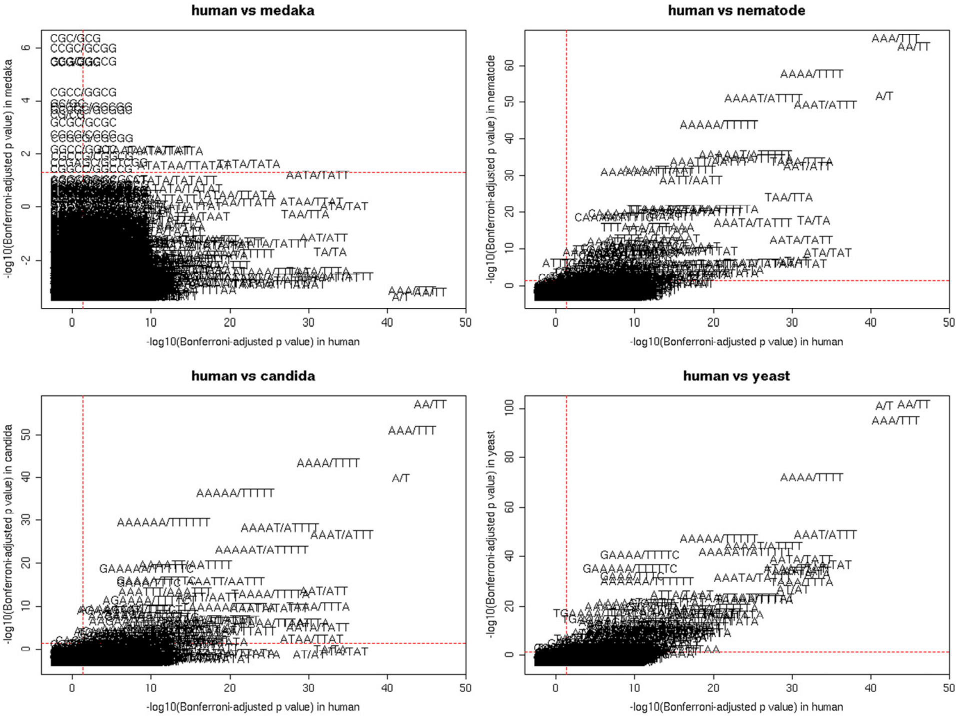

Figure 5: Comparative analyses of under-represented oligomers in nucleosomal DNA (over-represented in linker DNA). Significance level was set to 0.05 (indicated by dashed lines).

We also examined under-represented oligomers in nucleosomal DNA, which are interpreted as effective for nucleosome inhibition. Similar with over-represented motifs, the medaka nucleosomal DNA showed a clear difference with the other species (Figure 5 and Supplementary Table 3). GC-rich oligomers, such as CG/CG, CCG/CGG and CCGC/GCGG, were under-represented only in medaka. Although the nucleosome inhibition was previously observed with the consecutive runs of CG /CG [3], it was not enriched in the linker regions of the four species.

On the other hand, in our analysis, AT-rich oligomers were significantly responsible for the nucleosome inhibition, and each organism showed several specific under-represented motifs. The sequences associated with poly(A)/poly(T), such as AAAA/TTTT and AAAAT/ATTTT, were commonly observed in human, nematode, candida and yeast. It has been reported that the poly(A)/poly(T) is often observed in non-nucleosomal DNA of several organisms [2, 3] and that it can influence the disruption of nucleosomes by forming a specific DNA structure [21]. In addition, oligomers similar to the TATA box, such as TATA/TATA and TAATA/TATTA, were strongly frequent in the linker regions in human, medaka, nematode and yeast, but were not significant in candida ($P > 0.05$). The contribution of TATA box to nucleosome formation has been argued in several previous analyses. While Widlund *et al.* showed that TATA-containing sequences form stable nucleosomes by an *in vitro* assay [22], Ioshikhes *et al.* showed that nucleosomes in TATA-containing promoters are fuzzier than in TATA-less promoters [5].

The results of over- and under-represented oligomers in nucleosomal DNA suggest that each organism has somewhat different sequence tendency for nucleosome formation and avoidance. Several reported motifs were not always significant in our test based on the *in vivo* data. It is likely that the rather varied performance of inter-species predictions is due to these different sequence features.

Our approach has a limitation that the performance of each algorithm is evaluated assuming fixed nucleosome locations. It is known that some nucleosomes move in response to external stimuli. Since our evaluation dataset does not contain such information, we cannot evaluate whether the score produced by each algorithms reflects the stability of nucleosome positioning. In addition, repetitive elements were not evaluated in this study in which nucleosomal and linker DNA sequences overlapping repetitive elements were removed. However, several repetitive elements are strongly associated with nucleosome positioning (Tanaka, Yamashita, and Nakai, submitted).

4 Conclusions

As far as we know, it is the first third-party assessment of publicly available tools for the prediction of nucleosome positioning. From the AUC criterion, we suggest that Gupta's SVM with the RBF kernel is the best predictor. However, this method currently requires the training of a model with a number of nucleosomal and non-nucleosomal DNAs and deterioration of prediction accuracy when applying a SVM trained by the data of a different organism may occur. Therefore, for the prediction of a species' data without appropriate samples, Segal's method may be recommended because it does not require the training by users and shows relatively stable prediction accuracy in four species.

We hope that our results will be useful not only for practical purposes but also for the understanding of how much the nucleosome positioning is dependent on the local sequence features and how much the sequence determinants are common between species. Since many prediction tools are frequently updated or newly developed, the assessment should be updated periodically.

Additional Data and URL

The details of how to construct the genome-wide nucleosome maps are also described in the website (http://www.hgc.jp/~ytanaka/assess2009/index.html). Our test dataset and supplementary tables are also available in this site.

Acknowledgments

We thank all lab members for valuable advices and discussions. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was supported in part by the Global COE Program (Center of Education and Research for Advanced Genome-Based Medicine), MEXT, Japan.

References

[1] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P. and Widom, J. A genomic code for nucleosome positioning. *Nature* 442:772-778,2006.

[2] Satchwell, S. C., Drew, H. R. and Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 191:659-675,1986.

[3] Morohashi, N., Yamamoto, Y., Kuwana, S., Morita, W., Shindo, H., Mitchell, A. P. and Shimizu, M. Effect of sequence-directed nucleosome disruption on cell-type-specific repression by alpha2/Mcm1 in the yeast genome. *Eukaryot Cell* 5:1925-1933,2006.

[4] Wong, B., Chen, S., Kwon, J. A. and Rich, A. Characterization of Z-DNA as a nucleosome-boundary element in yeast *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 104:2229-2234,2007.

[5] Ioshikhes, I. P., Albert, I., Zanton, S. J. and Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nat Genet* 38:1210-1215,2006.

[6] Miele, V., Vaillant, C., d'Aubenton-Carafa, Y., Thermes, C. and Grange, T. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res* 36:3746-3756,2008.

[7] Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K. and Weng, Z. Nucleosome positioning signals in genomic DNA. *Genome Res* 17:1170-1177,2007.

[8] Yuan, G. C. and Liu, J. S. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol* 4:e13,2008.

[9] Tolstorukov, M. Y., Choudhary, V., Olson, W. K., Zhurkin, V. B. and Park, P. J. nuScore: a web-interface for nucleosome positioning predictions. *Bioinformatics* 24:1456-1458,2008.

[10] Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A. and Noble, W. S. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* 4:e1000134,2008.

[11] Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J. and Segal, E. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* 4:e1000216,2008.

[12] Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J. and Segal, E. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362-366,2009.

[13] Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., Wei, G. and Zhao, K. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132:887-898,2008.

[14] Sasaki, S., Mello, C. C., Shimada, A., Nakatani, Y., Hashimoto, S., Ogawa, M., Matsushima, K., Gu, S. G., Kasahara, M., Ahsan, B., Sasaki, A., Saito, T., Suzuki, Y., Sugano, S., Kohara, Y., Takeda, H., Fire, A. and Morishita, S. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* 323:401-404,2009.

[15] Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A. and Johnson, S. M. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* 18:1051-1063,2008.

[16] Jiang, H. and Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24:2395-2396,2008.

[17] Field, Y., Fondufe-Mittendorf, Y., Moore, I. K., Mieczkowski, P., Kaplan, N., Lubling, Y., Lieb, J. D., Widom, J. and Segal, E. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat Genet* 41:438-445,2009.

[18] Ozsolak, F., Song, J. S., Liu, X. S. and Fisher, D. E. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* 25:244-248,2007.

[19] Bharanidharan, D. and Gautham, N. Principal component analysis of DNA oligonucleotide structural data. *Biochem Biophys Res Commun* 340:1229-1237,2006.

[20] Richmond, T. J. and Davey, C. A. The structure of DNA in the nucleosome core. *Nature* 423:145-150,2003.

[21] Shimizu, M., Mori, T., Sakurai, T. and Shindo, H. Destabilization of nucleosomes by an unusual DNA conformation adopted by poly(dA) small middle dotpoly(dT) tracts *in vivo*. *EMBO J* 19:3358-3365,2000.

[22] Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E., Kahn, J. D., Crothers, D. M. and Kubista, M. Identification and characterization of genomic nucleosome-positioning sequences. *J Mol Biol* 267:807-817,1997.