MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **68** (2012) 639-650

ISSN 0340 - 6253

Prediction of Nucleosome Positioning Using the Dinucleotide Absolute Frequency of DNA Fragment

Zhiqian Zhang,[1] Yusen Zhang,[1*] Wei Chen,[1]
Ivan Gutman,[2] Yuchun Li[3]

[1] School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China
[2] Faculty of Science, University of Kragujevac, P. O. Box 60, 34000 Kragujevac, Serbia
[3] Marine College, Shandong University at Weihai, Weihai 264209, China

(Received January 29, 2012)

Abstract. Nucleosome is the basic structure of chromatin in eukaryotic cells, forming the chromatin fiber interconnected by sections of linker DNA. Nucleosome positioning is of great significance for the regulation of gene transcription. A few computational models have been proposed to predict in vivo nucleosome positioning on genome directly from DNA sequences. These approaches which vary from likelihood models to comparative genomics and supervised learning strategies, achieved limited success in prediction of nucleosome positions. Thus, development of new computational methods based on multiple factors is desirable. We use a support vector machine (SVM) with the absolute frequency of DNA fragments to predict nucleosomal DNA sequences. Computational experiments on several nucleosome positioning datasets show that in all cases the proposed model gives better prediction performance than other models. The results of this study have important implications for models of sequence-dependent positioning, since they suggest that a dinucleotide absolute frequency is involved in preferred nucleosome occupancy. So, the model is useful for predicting nucleosome positioning.

1 Introduction

DNA in eukaryotic nuclei is assembled into chromatin. The primary function of chromatin is to compact genomic DNA, that otherwise would not fit into the cell nucleus. Nucleosomes are the basic repeating units of chromatin. They are composed of octamers of histone proteins

*Corresponding author: zhangys@sdu.edu.cn

around which about 147 base pairs of DNA are tightly wrapped. Nucleosomes are separated by linker DNA that may vary in length and provides flexibility to the nucleosome chain. The precise location of the nucleosome core DNA's in genomic DNA is the nucleosome's positioning, playing an important role in many biological processes, including replication, transcription, DNA repair, etc [1] [2][3] [4]. How the DNA sequence and deformation affect the positioning of nucleosomes has been recently the subject of extensive coverage [5] [6] [7] [8] [9]. Genomic DNA sequences possess high variability in their binding affinity to the nucleosome core.

Although numerous factors can contribute towards determining the nucleosome positioning in vivo, it is widely accepted that the specificity of the interaction of the core histone octamer with respect to the underlying DNA sequence plays a important role [10] [11]. Defining the sequence properties of nucleosomal DNA has been the topic of a number of studies, with focus on the prediction of nucleosome positioning. The identification of patterns in the occurrence and distribution of short nucleotide motifs in nucleosomal sequences provides an insight into how the DNA sequence or structure may determine nucleosome positioning [12] [13] [14].

Several studies predicted in vivo nucleosome positions directly from the nucleosomes' intrinsic DNA sequence preferences, that vary greatly between differing DNA sequences. Segal et al isolated nucleosome–bound sequences at high resolution from yeast and used these sequences to construct a dinucleotide–based model for nucleosome positioning [12]. Ioshikhes et al. utilized the occurrence of periodically distributed AA and TT dinucleotides to define a 'nucleosome positioning sequence' [13]. Both groups subsequently applied their models to predict nucleosome positioning on the entire *Saccharomyces cerevisiae* genome and compared their predictions to experimentally determined nucleosome locations. The results obtained suggest that the genome DNA sequence partly determines the locations of nucleosomes. Yuan et al. defined the nucleosomal positions for a significant portion of the yeast genome using micrococcal nuclease digestion followed by microarray hybridization [20]. Lee et al. used the same technique at a higher resolution to map nucleosomes along the entire yeast genome [21], and Shivaswamy et al. defined the yeast nucleosome positions under two different conditions using high–throughput sequencing [22]. Field et al. and Kaplan et al. devised a computational model in which nucleosome occupancy is governed

only by the intrinsic sequence preferences of nucleosomes [17] [14]. Moreover, since in vitro nucleosome depletion is evident at many transcription factor binding sites and around gene start and end sites, they suggest that also nucleosome depletion at these sites in vivo is partly encoded in the genome. Yuan et al. proposed an N-score model to discriminate nucleosome and linker DNAs using wavelet energies as covariates in a logistic regression model [23]. In the same year, a web–interface called 'nuScore' was developed for estimating the affinity of histone core to DNA and prediction of nucleosome positioning, based on the DNA deformation energy score[24]. Using a nucleosome DNA sequence probe based on a specific dinucleotide periodical pattern, Salih et al. introduced a straightforward method for nucleosome mapping [25]. Albert et al. sequenced DNA from 322,000 individual *S. cerevisiae* nucleosomes and analyzed the functions of nucleosome positioning in gene regulation [26]. Peckham et al. trained a support vector machine (SVM) with the frequency of k-mer (k = 1 to 6) DNA fragments in nucleosomal DNA sequences to predict nucleosome positions [11]. Ogawa et al. refined the method described by Peckham et al. and suggested that TGG/CCA and CAG/CTG were key sequence fragments in the formation of nucleosomes [34]. Existing methods extract specific dinucleotide sequence patterns from nucleosomal sequences, and used them to predict nucleosome positioning sequences[12] [28]. Regardless of the computational approaches undertaken, which vary from likelihood models to comparative genomics and supervised learning strategies, their accuracy remains limited. Overall, the observations made by various research groups using both experimental and theoretical approaches imply that there are few regions in the genome in which the nucleosomal landscape is consistent across the cellular population, and that the majority of the nucleosomes are stochastically positioned. The primary sequence conservation seems to be of little assistance in attempts to define the population of consistently positioned nucleosomes [28].

Based on the absolute frequency of dinucleotide of DNA sequence [29] [30], we developed a new method for predicting nucleosome positioning from genome sequences. Our model for distinguishing nucleosome and linker DNAs in yeast, human, medaka, nematode, and candida genomes has better performance than previous works, with both high predictive success rates and simpler computation.

2 Methods

2.1 Nucleosome positioning data

The nucleosome maps are obtained from two articles [31][22]. The datasets of Whitehouse et al. and Shivaswamy et al. are used for analysis and comparison of common characteristics of nucleosomal DNA sequences. The data of human, medaka, nematode, candida, and yeast from Tanaka et al.[32], available at http://www.hgc.jp/ ytanaka/assess2009/index.html, are used for further validation of prediction model. For each organism, the data includes 10 evaluation datasets with randomly extracted 100 nucleosomal and 100 linker DNA sequences in each.

2.2 A model for discriminating nucleosome forming and inhibiting sequences

The authors find that the dinucleotide absolute frequency can reveal other characteristics of the nucleosome forming sequence. Consider a DNA sequence read from the 5'- to the 3'-end with n bases. The occurrences of the nucleotide X (A, C, G, or T), is denoted by the positive integer X_n. By considering two neighboring bases, we can obtain sixteen dinucleotides XY: AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG and CC. The occurrences of the dinucleotide XY is denoted by the positive integer XY_n. For any DNA sequence f, the dinucleotide absolute frequency $P_f(XY)$ is defined as the ratio of total occurrences of the dinucleotide XY to that of the first nucleotide X composing this dinucleotide. That is

$$P_f(XY) = XY_n/X_n \ . \tag{1}$$

We construct a 16-component vector:

$$x(f) = \Big(P_f(AA), P_f(AT), ..., , P_f(CG), P_f(CC)\Big)$$

and then establish the correspondence between the DNA sequence and $x(f)$. However, the lengths of nucleosomes on different chromosomes are different, even within the same chromosome. In order to eliminate the impact of these length differences, we deal with the following vector:

$$X(f) = \frac{147}{n}\, x(f) = \frac{147}{n}\Big(P_f(AA), P_f(AT), \ldots, P_f(CC)\Big) \ .$$

In this way, each DNA sequence is represented as a 16-component vector, in which each entry is a normalized absolute frequency of a particular dinucleotide. These vectors were used to train LIBSVM, which is a publicly available online library for training and predicting with SVM [33]. As a supervised machine learning technology, it has been successfully used in wide areas of bioinformatics by transforming the input vector into a high-dimension Hilbert space and to seek a separating hyperplane in this space. For a two–class classification problem, a series of training vectors were marked by +1 and -1, which respectively indicate the two classes. After training, predictions can be made by predicting the associated +1/ -1 label for each test sample. When using LIBSVM, it is important to correctly choose the parameters c and g. In this work, we set $c = 4$ and $g = 2$.

2.3 Evaluation of prediction performance

In order to evaluate the performance of a model, selecting a test method is an important issue. In the previous papers, the jackknife test and ROC curve were used normally. The ROC (Relative Operating Characteristic curve), is a comparison of two operating characteristics (TP & FP) as the criterion changes. AUC (the area under the ROC curve) is also used to evaluate performance of model. The AUC provides a single measure of overall prediction accuracy. The values 0.5 of AUC is equivalent to random prediction. Values of AUC between 0.5 and 0.7 indicate poor accuracy. Values of AUC between 0.7 and 0.9 indicate good prediction accuracy and above 0.9 indicate excellent prediction accuracy. The overall prediction accuracy (A) of the five models is defined as

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

whereas sensitivity S, specificity P, and Matthew's correlation coefficient MCC are defined as

$$S = \frac{TP}{TP + FN}, \tag{3}$$

$$P = \frac{TP}{TP + FP}, \tag{4}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{5}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

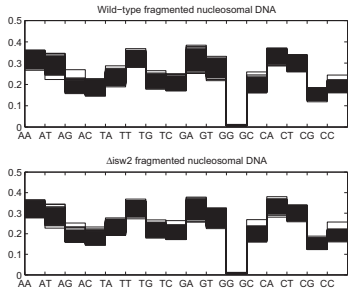

Figure 1: Stairstep plot of the dinucleotide absolute frequency for data of Whitehouse et al.

3 Results and discussion

3.1 Analysis of common characteristics

Although numerous factors can contribute towards determining nucleosome positioning, it is widely accepted that the specificity of the interaction of the core histone octamer with respect to the underlying DNA sequence plays a major role [14] [12] [13], and the other papers suggested that genomic DNA play an important role in nucleosome positioning, the identification of the patterns in the occurrence and distribution of short nucleotide motifs in nucleosomal sequences provides an insight into how DNA sequence or structure determines nucleosome positioning and is used to predict histone octamer positioning from DNA sequence [11] [34] [12] [28]. But several recent papers reported the opposite opinion (genomic DNA is not a major factor) [15]. These facts suggest that there may be additional sequence–independent signals that are important for nucleosome positioning, undiscovered by current models.

Factors other than the surrounding DNA sequence might contribute to nucleosome positioning in vivo. For example, nucleosome remodelling complexes, such as Isw2 in yeast, override the sequence preferences of nucleosomes, causing intrinsic nucleosome eviction or repositioning of the nucleosome away from the unfavourable sequence [35]. There are several studies of nucleosome positions in yeast whose main focus is not on intrinsic sequence preferences nor on the nucleosome organization with respect to various genomic features, but

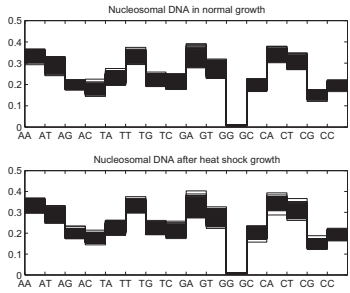

Figure 2: Stairstep plot of the dinucleotide absolute frequency for data of Shivaswamy et al.

rather on how chromatin responds either to environmental perturbations or to deleting genes implicated in chromatin remodeling and maintenance [36].

The authors [31] investigated the role of the ATP-dependent chromatin-remodeling complex Isw2 in controlling chromatin structure across the yeast genome and sought to discover Isw2 targets genome-wide by identifying differences in nucleosome positions between wild-type and Δisw2 mutant strains, concluded that Isw2 functions by moving nucleosomes toward intergenic regions, where many important regulatory sequences are located. The ability of Isw2 and other chromatin-remodeling enzymes to actively reposition nucleosomes demonstrates that intrinsic nucleosome-positioning preferences may be disrupted in living cells. In order to identify the intrinsic characteristics that could determine nucleosome positioning, we analyzed the publicly available sequencing data of Whitehouse et al. These data are in turn divided into different groups with 100 per group and the 100-mean dinucleotide absolute frequency for each group was calculated by averaging the individual dinucleotide absolute frequencies over all the DNA fragments in the group. The profiles of 100-mean dinucleotide absolute frequency of 63000 (19 nucleosomal DNA sequences excluded) wild-type and 62500 (86 nucleosomal DNA sequences excluded) Δisw2 fragmented nucleosomal DNA sequence reads mapped on the genome are calculated and shown in Fig. 1, respectively. Although the ATP-dependent chromatin-remodeling complex Isw2 can catalyses the directional shift of nucleosomes towards intergenic regions [31], our data illustrate that Δisw2 fragmented nucleosomal DNA sequence reads are highly similar with wild-type ones.

Table 1: Comparison of AUC values for different models

	human	medaka	nematode	candida	yeast	average
Segal (ver.3)	0.694	0.516	0.708	0.722	0.764	0.681
Segal (ver.2)	0.684	0.53	0.717	0.752	0.804	0.697
Segal (ver.1)	0.487	0.565	0.492	0.51	0.514	0.514
Miele	0.333	0.508	0.319	0.425	0.313	0.379
Gupta (Linear)	0.611	0.605	0.696	0.678	0.802	0.678
Gupta (Quadratic)	0.611	0.605	0.697	0.682	0.794	0.678
Gupta (Cubic)	0.596	0.634	0.702	0.673	0.799	0.681
Gupta (RBF1)	0.695	0.705	0.743	0.69	0.811	0.729
Gupta (RBF5)	0.641	0.659	0.744	0.703	0.796	0.709
Gupta (RBF10)	0.657	0.642	0.736	0.705	0.798	0.707
Our model	0.872	0.884	0.836	0.766	0.831	0.838

In order to know how chromatin structure in yeast cells responds to physiological perturbations such as heat shock that are usually accompanied by massive transcriptional changes. The authors [22] subjected yeast cells grown in a rich medium to a 15-min period of heat shock and generated a differential map of nucleosome positions, which consisted of 514,803 and 1,036,704 uniquely mapped reads for the normal and heat-shock growth conditions, respectively. As before, the profiles of 100-mean dinucleotide absolute frequency of the 49000 (43 nucleosomal DNA sequences excluded) normal growth and 52800 (17 nucleosomal DNA sequences excluded) heat-shock growth higher confidence nucleosomal DNA sequences are calculated and shown in Fig. 2, respectively. Comparison between the 100-mean dinucleotide absolute frequency distribution of nucleosomal DNA sequences under normal growth and heat-shock growth conditions reveals a striking overall similarity and further demonstrate chromatin structure is largely invariant with respect to different growth conditions.

These comparisons revealed striking similarities in profiles of dinucleotide absolute frequency of nucleosomal DNA sequences in different dataset under different conditions, suggesting that the dinucleotide absolute frequency distribution of DNA sequence also is an important factor in the control of nucleosome positioning.

We repeated the SVM cross–validation testing procedure, using data generated by Tanaka et al. for human, medaka, nematode, candida, and yeast [32]. In 10-fold cross–validation, the positive dataset and the negative dataset were divided at random into ten subsets for each

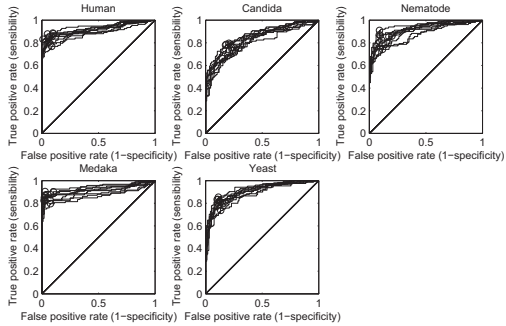

Figure 3: ROC curves of 10-fold test on the data of five organisms

of the five organisms: positive training set (90% of the positive dataset data) and positive test set (the left-out data), negative training set (90% of the negative dataset data) and negative test set (the left-out data), respectively. The positive and negative training sets form the training set, The positive and negative test sets form the test set. In the training set, every sequence in the positive training set is marked by 1, and every sequence in the negative training set by -1. By this a mark vector is obtained. These vectors were used to train a support vector machine. After training a support vector machine, the test set and its mark vector were repeated ten times using a different leave-out set each time. The ROC of these data is shown in the Figs. 3 and the AUC values of all prediction methods applied to five organisms are listed in Table 1. Compared with the AUC values of human, medaka, nematode, candida, and yeast reported in [32], we find that our method is more accurate.

4 Conclusion

We have developed a novel computational approach for the prediction of nucleosome positioning. We find that our model has a significantly improved performance relative to the previous models with regard to the ability to recognize known nucleosome and linker DNA sequences. Our use of dinucleotide composition distributions is motivated by the fact that these are the simplest elements that can capture the sequence–dependence of DNA bending. Our results showed dinucleotide absolute frequency of sequence can be used to predict

nucleosome sequences and that utilization of the SVM can improve the accuracy of these predictions.

Acknowledgement. This study was supported by the Shandong Natural Science Foundation (Grant No. ZR2010AM020), the Serbian Ministry of Science and Education (Grant No. 174033), and the Major International (Regional) Joint Research Project of NSFC (Grant No. 31110103910).

References

[1] R. T. Simpson, D. W. Stafford, Structural features of a phased nucleosome core particle, *Proc. Natl. Acad. Sci. USA* **80** (1983) 51–55.

[2] M. Kato, Y. Onishi, Y. Wada–Kiyama, T. Abe, T. Ikemura, S. Kogan, A. Bolshoy, E. N. Trifonov, R. Kiyama, Dinucleosome DNA of human K562 cells: experimental and computational characterizations, *J. Mol. Biol.* **60** (1990) 719–731.

[3] P. T. Lowary, J. Widom, New DNA sequence rules for high affinity binding to histone octamer and sequence–directed nucleosome positioning, *J. Mol. Biol.* (1998) **276** 19–42.

[4] A. Flaus, K. Luger, S. Tan, T. J. Richmond, Mapping nucleosome position at single base–pair resolution by using site–directed hydroxyl radicals, *Proc. Natl. Acad. Sci. USA* **93** (1996) 1370–1375.

[5] A. Travers, E. Hiriart, M. Churcher, M. Caserta, E. D. Mauro, The DNA sequence–dependence of nucleosome positioning in vivo and in vitro, *J. Biomol. Struct. Dyn.* **27** (2010) 713–724.

[6] F. Xu, W. K. Olson, DNA architecture, deformability, and nucleosome positioning, *J. Biomol. Struct. Dyn.* **27** (2010) 725–739.

[7] E. N. Trifonov, Nucleosome positioning by sequence, state of the art and apparent finale, *J. Biomol. Struct. Dyn.* (2010) **27** 741–746.

[8] D. J. Clark, Nucleosome positioning, nucleosome spacing and the nucleosome code, *J. Biomol. Struct. Dyn.* **27** (2010) 781–793.

[9] G. Arya, A. Maitra, S. A. Grigoryev, A structural perspective on the where, how, why, and what of nucleosome positioning, *J. Biomol. Struct. Dyn.* **27** (2010) 803–820.

[10] E. N. Trifonov, Sequence–dependent deformational anisotropy of chromatin DNA, *Nucleic Acids Res.* **8** (1980) 4041–4053.

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

[12] E. Segal, Y. Fondufe–Mittendorf, L. Chen, A genomic code for nucleosome positioning, *Proc. Natl. Acad. Sci. USA* **442** (2006) 772–778.

[13] I. P. Ioshikhes, I. Albert, S. J. Zanton, B. F. Pugh, Nucleosome positions predicted through comparative genomics, *Nat. Genet.* **38** (2006) 1210–1215.

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

[15] Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, K. Struhl, Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo, *Nat. Struct. Mol. Biol.* **16** (2009) 847–852.

[16] H. R. Drew, A. A. Travers, DNA bending and its relation to nucleosome positioning, *J. Mol. Biol.* **186** (1985) 773–790.

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

[18] T. Bettecken, E. Trifonov, Repertoires of the nucleosome–positioning dinucleotides, *PLoS ONE* **4** (2009) e7654.

[19] M. Kato, Y. Onishi, Y. Wada–Kiyama, T. Abe, T. Ikemura, S. Kogan, A. Bolshoy, E. N. Trifonov, R. Kiyama, Dinucleosome DNA of human K562 cells: experimental and computational characterizations, *J. Mol. Biol.* (2003) **332** 111–125.

[20] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, O. J. Rando, Genome–scale identification of nucleosome positions in *S. cerevisiae Science*, **309** (2005) 626–630.

[21] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, A high–resolution atlas of nucleosome occupancy in yeast, *Nat. Genet.* **39** (2007) 1235–1244.

[22] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, V. R. Iyer, Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLoS Biology* **6** (2008) e65.

[23] G. C. Yuan, J. S. Liu, Genomic sequence is highly predictive of local nucleosome depletion, *PLoS Comput. Biol.* **4** (2008) 164–174.

[24] M. Tolstorukov, V. Choudhary, W. Olson, V. Zhurkin, P. Park, NuScore: a web-interface for nucleosome positioning predictions, *Bioinformatics* **28** (2008) 1456–1458.

[25] F. Salih, B. Salih, E. N. Trifonov, Sequence–directed mapping of nucleosome positions, *J. Biomol Struct. Dyn.* **24** (2007) 429–514.

[26] I. Albert, T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, B. F. Pugh, Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome, *Nature* **446** (2003) 572–576.

[27] R. Ogawa, N. Kitagawa, H. Ashida, R. Saito, M. Tomita, Computational prediction of nucleosome positioning by calculating the relative fragment frequency index of nucleosomal sequences, *FEBS Lett.* **584** (2010) 1498–1502.

[28] N. Christoforos, S. Althammer, M. Beato, R. Guig, Structural constraints revealed in consistent nucleosome positions in the genome of S. cerevisiae, *Epigenetics Chromatin* (2010) **3** 20.

[29] Y. S. Zhang, W. Chen, A dissimilarity measure based on free energy of DNA nearest–neighbor interaction, *J. Biomol. Struct. Dyn.* **28** (2011) 557–565.

[30] Y. S. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 477–488.

[31] I. Whitehouse, O. J. Rando, J. Delrow, T. Tsukiyama, Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450** (2007) 1031–1035.

[32] Y. Tanaka, K. Nakai, An assessment of prediction algorithms for nucleosome positioning, *Genome Inf.* **23** (2009) 169–178.

[33] C. Chih-Chung, L. Chih-Jen, LIBSVM, a library for support vector machines (2001). Available at http://www.csie.ntu.edu.tw/cjlin/libsvm .

[34] R. Ogawa, N. Kitagawa, H. Ashida, R. Saito, M. Tomita, Computational prediction of nucleosome positioning by calculating the relative fragment frequency index of nucleosomal sequences, *FEBS Lett.* **584** (2010) 1498–1502.

[35] C. Jiang, B. F. Pugh, Nucleosome positioning and gene regulation: advances through genomics, *Genetics* **10** (2009) 161–172.

[36] D. Tolkunov, A. V. Morozov, Genomic studies and computational predictions of nucleosome positions and formation energies, *Adv. Protein Chem. Struct. Biol.* **79** (2010) 1–57.