MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **72** (2014) 551-562

ISSN 0340 - 6253

Nucleosome Positioning Based on Information Content

Yan Su[a], **Yusen Zhang**,[a,1] **Wei Chen**[a],
Ivan Gutman[b]

[a]*School of Mathematics and Statistics, Shandong University at Weihai,
Weihai 264209, China*

[b]*Faculty of Science, University of Kragujevac,
P. O. Box 60, 34000 Kragujevac, Serbia*

(Received May 19, 2014)

Abstract

Nucleosome is the basic structure of chromatin in eukaryotic cells. Nucleosome positioning plays a key role in the regulation of many biological processes like replication, transcription and DNA repair. In this paper the informational entropy and the mutual information are applied to detect the information on nucleotide correlation stored in the nucleosomal sequences. We find that the two nucleotides separated by a gap of length 1 have higher certainty than in the case of longer gaps. Also, two nucleotides separated by a gap of length 1,2 have a much higher correlation, compared to longer gaps. This finding was used to construct a feature vector suitable fort classifying nucleosomal and linker sequences. Computational experiments on several nucleosome positioning datasets show that in all cases the proposed model gives a better prediction performance than other models. This suggests that our vector contains important signaturs of nucleosome positioning.

1 Introduction

A chromosome is an organized structure of DNA and protein found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Nucleosome is the basic structure of chromatin in eukaryotic cells, forming the chromatin fiber, interconnected by sections of linker DNA. Each nucleosome contains

[1]Corresponding author: zhangys@sdu.edu.cn

approximately 165 bp genomic DNA, and the core nucleosome is about 147 bp genomic DNA, wrapped in 1.75 turns around an octamer of the histone proteins H2A, H2B, H3, and H4 (see [1]). Neighboring nucleosomes are separated from each other by 10-50-bp-long stretches of unwrapped linker DNA [2]. Thus 75 to 90 percent of eukaryotic genomic DNA is packaged in nucleosomes. The precise location of the nucleosome core DNA's in genomic DNA is the nucleosome's positioning, playing an important role in replication, transcription, DNA repair, and many other biological processes [3–6].

Numerous factors can contribute towards determining the nucleosome positioning in vivo, and many studies have provided extensive evidence indicating a sequence dependent positioning of nucleosomes along DNA [7, 8]. The CA dinucleotide has been shown to be important for nucleosome positioning, and the decamer TATAAACGCC has a high affinity for histones [9, 10]. Poly (dA:dT) has been shown to increase the accessibility of transcription factors bound to nearby sequences [11]. Analysis of periodicities in genomic DNA is also important for clarifying the basic genomic structures. In previous works, Satchwell et al. [12] detected a periodicity of 10 bp in the chicken nucleosome sequence. There is evidence of a periodic repeating at every 10.4 bases of the dinucleotides AA and TT in nucleosome forming sequences [13]. Besides, several prediction tools have been developed for nucleosome positioning. Segal et al. in 2006 [14] used a hidden Markov model for constructing a "nucleosome-DNA interaction model". Their model has a 50% predicting accuracy. In 2008, Guo–Cheng Yuan et al. [15] proposed an N-score model fot discriminating nucleosome and linker DNAs, using wavelet energies as covariates in a logistic regression model. Peckham et al. [8] and Gupta et al. [16] introduced support vector machines (SVMs) to classify nucleosomal and non-nucleosomal DNAs. Using the frequencies of the k-mers for $k = 1$ to 6, Peckham et al. used a SVM to distinguish between nucleosome forming and nucleosome inhibiting sequences [8]. They found that the GC/AT richness of a sequence was the strongest single factor among k-mer frequencies in determining its nucleosome formation potential.

We first make a research of the informational entropy and the mutual information in the nucleosomal sequences. We find that the two nucleotides separated by a gap of length 1,2 have a high correlation compared to the others. Based on these results, a new method for predicting nucleosome positioning from genome sequences is developed. In the case of yeast, human, medaka, nematode, and candida, it has a better performance for distinguishing nucleosome and linker DNAs than the previous works.

2 Materials and Methods

2.1 Nucleosome positioning data

For the validation of the prediction model we employed the data of human, medaka, nematode, candida, and yeast from Tanaka et al. [17], available at `http://www.hgc.jp/ytanaka/assess2009/index.html`. For each organism, the data include 10 evaluation datasets with randomly extracted 100 nucleosomal and 100 linker DNA sequences.

2.2 Methods

In this work, an information–theoretic method is used to to describe the characteristics of the nucleosome sequences $N_1, N_2, N_3, \ldots, N_s$. To each sequence N_i , the informational entropy H_i and the mutual information $I_{k,i}$ of every position are calculated as

$$H_i = -\sum_{x \in \alpha} p_{(x)} \log p_{(x)}$$

and

$$I_{k,i} = \sum_{x \in \alpha, y \in \alpha} p_{x(k)y} \frac{\log p_{x(k)y}}{p_{(x)} p_{(y)}}$$

for $i = 1, 2, 3, \ldots, ls - k$ and $k = 1, 2, 3, \ldots$. Here $\alpha = \{A, G, C, T\}$ is the set of the nucleotides, ls is the length of the sequence, $p_{(x)}$ is the probability of finding the nucleotide x in the corresponding position of the sequence, and $p_{x(k)y}$ denotes the joint probability of finding the pair of nucleotides x and y, separated by a gap of length k.

The informational entropy H_i shows the uncertainty of the nucleotides in position i for one organism's sequences. $H_i = 0$ denotes an invariable nucleotide appearing in position i.

The mutual information $I_{i,k}$ measures the amount of information that can be obtained in position i from nucleotide x about another nucleotide y occurring at distance k after x. In a perfect dataset, uncorrelated pairs of not co-evolving positions $(i, i + k)$ would show $I_{i,k} = 0$ as the two-point probability is equal to the product of the one-point probabilities, $p_{x(k)y} = p_{(x)} \cdot p_{(y)}$, due to the independence of positions.

2.2.1 Support vector machine

The vectors considered below were used to train LIBSVM, which is a publicly available online library for training and predicting with SVM [18]. 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 seeking a separating hyperplane in this space. For a two–class classfication 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 $g = 2$.

2.2.2 Evaluation of prediction performance

For evaluating the performance of a model, the selection of a test method is an important issue. In previous papers, the jackknife test and ROC curve were used. 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 the model.

AUC provides a single measure of overall prediction accuracy. The 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

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

whereas the sensitivity S, specificity P, and Matthew's correlation coefficient MCC of every subcellular location are defined as

$$S = \frac{TP}{TP + FN} \qquad , \qquad P = \frac{TP}{TP + FP}$$

and

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

with TP, TN, FP, and FN representing true positive, true negative, false positive, and false negative, respectively.

3 Results and discussion

3.1 Informational entropy

For a particular organism, if we get the DNA sequences $N_1, N_2, N_3, \ldots, N_s$, then we can calculate the informational entropy $H_{k,i}$ of the two nucleotides, separated by a gap of length k in position i , $i = 1, 2, 3, \ldots, ls - k$, for $k = 1, 2, 3, \ldots$, where ls is the length of the underlying sequence. The expression for $H_{k,i}$ reads:

$$H_{k,i} = - \sum_{x \in \alpha, y \in \alpha} p_{x(k)y} \log p_{x(k)y}$$

where the set α and the probability $p_{x(k)y}$ are same as specified above.

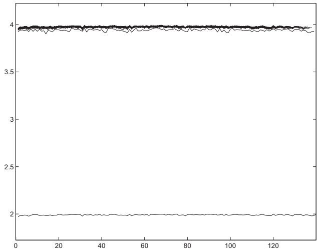


Fig. 1. The $H_{k,i}$-values of nucleosomal sequences. The x-axis represents the position in the sequences, the y-axis represents the informational entropy of the corresponding position in the sequences. The bottom line corresponds to $H_{1,i}$.

For medaka, we calculated $H_{k,i}$ for the nucleosomal sequences and the results are shown in Fig. 1. In order to see the difference clearly, in Fig. 2 we show the average informational entropy H_k traversing $ls - k$ positions, where

$$H_k = \frac{1}{ls - k} \sum_{i=1}^{ls-k} H_{k,i} \ .$$

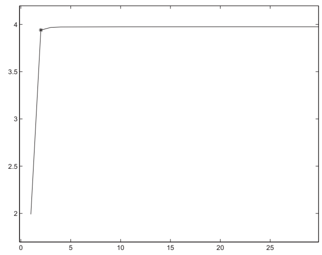


Fig. 2. The H_k-values of nucleosomal sequences. The x-axis represents the gap of the two nucleotides, the y-axis represents the average informational entropy traversing $ls - k$ positions.

From Fig. 2 we see that for $k \geq 2$, H_k does not significantly change, and that H_1 has the minimum value. This means that the dinucleotides $AA, AT, AG, AC, \ldots, CC$ have the smallest uncertainties, and this gives evidence for the sequence–dependent positioning of nucleosomes along DNA [7,8]. The results are consistent with the earlier conclusions that the dinucleotides AA and TT repeat at every 10.4 bases in nucleosome–forming sequences [12]. In addition, the CA dinucleotide was shown to be important for nucleosome positioning [9,10]. The low uncertainty of the dinucleotides $AA, AT, AG, AC, \ldots, CC$ also explains why the previous nucleosome positioning methods, using their frequencies, have a better performance [19].

3.2 The mutual information

In order to facilitate the observation, we calculate the average mutual information I_k as shown in Fig. 3, where

$$I_k = \frac{1}{ls - k} \sum_{i=1}^{ls-k} I_{k,i} \ .$$

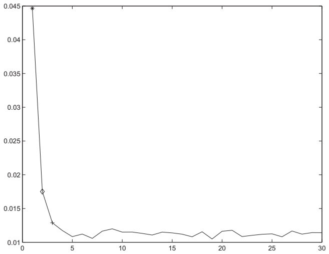


Fig. 3. The I_k-values of nucleosomal sequences. The x-axis represents the gap of the two nucleotides, the y-axis represents the average mutual information traversing $ls - k$ positions. We see that with the increase of the gap, I_k trends down. When $k \geq 3$, the values of I_k do not change significantly.

We calculated I_k of the nucleosome sequences of five organisms, see Table 1. From Table 1, we see that for all organisms the trend of I_k is roughly the same, although there are slight discrepancies for different organisms. The I_1-value is the largest, which is consistent with the the results on informational entropy outlined above. This, again means that two nucleotides separated by a gap of length 1 have the highest degree of correlation. The 3-bp periodicity associated with the codon usage largely creates the maximum of I_2 . The data in Table 1 indicate that when $k \geq 3$, then the values of I_k are quite small and thus the correlation is weak. Therefore, we use only I_1 and I_2 in our nucleosome positioning.

Table 1: The value of I_k of nucleosome sequences of five organisms

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
human	0.0580	0.0182	0.0140	0.0142	0.0123	0.0141	0.0137	0.0133
medaka	0.0447	0.0175	0.0129	0.0117	0.0108	0.0112	0.0106	0.0116
nematode	0.0400	0.0172	0.0146	0.0109	0.0113	0.0125	0.0111	0.0120
candida	0.0310	0.0154	0.0173	0.0126	0.0113	0.0173	0.0119	0.0119
yeast	0.0225	0.0128	0.0136	0.0114	0.0112	0.0114	0.0111	0.0109
average	0.0392	0.0162	0.0145	0.0122	0.0114	0.0133	0.0117	0.0119

3.3 Classification of nucleosomal and linker sequences

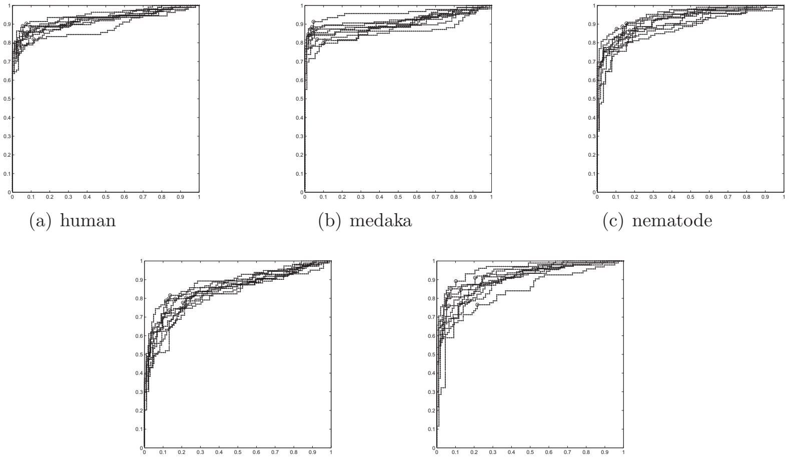
For each sequence, based on the above results, we get our brief feature vector $\mathbf{F} = (F_1, F_2)$ where

$$F_{k,x,y} = \frac{147 \cdot 4^{k-1} \cdot p_{x(k)y}}{\ell} \quad ; \quad k = 1, 2$$

and $x \in \alpha$, $y \in \alpha$, $\alpha = \{A, G, C, T\}$. In the above formula, ℓ is the length of the sequence. As we can see, $\mathbf{F}$ is a 32-dimensional vector.

We repeated the SVM cross–validation testing procedure, using data generated by Tanaka et al. For human, medaka, nematode, candida, and yeast (see Section 2). In 10-fold cross–validation, the positive dataset and the negative dataset were divided at random into ten subsets for each 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 the support vector machine, the test set and its mark vector were repeated ten times, each time using a different leave-out set. The ROC of these data is shown in Fig. 4 whereas the AUC-values of all prediction methods applied to five organisms are listed in Table 2.

Compared with the AUC-values of human, medaka, nematode, candida, and yeast, we find that our method is more accurate than previous researches. This shows that the way we structure the sequence can really infer significant features of the nucleosomal and linker DNA sequences.



(a) human (b) medaka (c) nematode

(d) candida (e) yeast

Fig. 4. ROC curves of five species.

Table 2: 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 (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
Zhang [20]	0.872	0.884	0.836	0.766	0.831	0.838
our model	0.9237	0.9068	0.9175	0.8482	0.9079	0.9008

4 Conclusion

The informational entropy and the mutual information are applied to detect the information on nucleotide correlation stored in the nucleosomal sequence. From the research of informational entropy, we learn that the two nucleotides separated by a gap of length 1 have the smallest uncertainty. This gives evidence for the sequence–dependent positioning of nucleosomes along DNA [7,8], and indicates the importance of the dinucleotide. Analysis of mutual information showed that the two nucleotides separated by a gap of length 1,2 have a high correlation, compared to the others. We used this finding to construct our feature vector for classifying the nucleosomal and linker sequences. We found (cf. Table 2) that our model has an improved performance relative to the previous models. This suggests that our vector contains important signatures of nucleosome positioning.

On the other hand, nucleosome positioning along genome is determined by multiple factors, including preference of DNA sequences, competitive or cooperative binding of protein factors, activities of ATP-dependent remodeling complexes, and so on [21–24]. If we add periodicity, curvature, or other factors to our vector, the results may become better. This is planned to be our next research. Characterizing the nucleosome positioning in the whole sequence is the main research object in the future.

Acknowledgements. This study was supported by the Shandong Natural Science Foundation (Grant No. ZR2010AM020) and the Major International (Regional) Joint Research Project of NSFC (Grant No. 31110103910).

References

[1] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, T. J. Richmond, *J. Mol. Biol.* **319** (2002) 1097–1113.

[2] K. E. van Holde, *Chromatin*, Springer, New York, 1989.

[3] P. T. Lowary, J. Widom, New DNA sequence rules for high afnity binding to histon octamer and sequence–directed nucleosome positioning, *J. Mol. Biol.* **276** (1998) 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] R. T. Simpson, D. W. Staord, Structural features of a phased nucleosome coreparticle, *Proc. Natl. Acad. Sci. USA* **80** (1983) 51–55.

[6] 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.

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

[8] 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.

[9] H. R. Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P. E. Nielsen, J. D. Kahn, D. M. Crothers, M. Kubista, *J. Mol. Chem.* **267** (1997) 807–817.

[10] H. R. Widlund, P. N. Kuduvalli, M. Bengtsson, H. Cao, T. D. Tullius, M. Kubista, *J. Mol. Chem.* **264** (1999) 31847–31852.

[11] K. Struhl, Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast, *Proc. Natl. Acad. Sci. USA* **82** (1985) 8419–8423.

[12] S. C. Satchwell, H. R. Drew, A. A. Travers, *J. Mol. Biol.* **191** (1986) 659–675.

[13] A. B. Cohanim, Y. Kashi, E. N. Trifonov, Yeast nucleosome DNA pattern: Deconvolution from genome sequences of S. cerevisiae, *J. Biomol. Struct. Dyn.* **22** (2005) 687–694.

[14] E. Segal, Y. Fondufe–Mittendorf, L. Chen, A. Thastrom, Y. Field, I. Moore, J. P. Wang, J. Widom, *Nature* **442** (2006) 772–778.

[15] G. C. Yuan, J. S. Liu, *PLoS Comput. Biol.* **4** (2008) e13.

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

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

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

[19] Z. Zhang, Y. Zhang, I. Gutman, Predicting nucleosome positions in yeast: Using the absolute frequency, *J. Biomol. Struct. Dyn.* **29** (2012) 1081–1088.

[20] Z. Zhang, Y. Zhang, W. Chen, I. Gutman, Y. Li, Prrediction of nucleosome positioning using the dinucleotide absolute frequency of DNA fragment, *MATCH Commun. Math. Comput. Chem.* **68** (2012) 639–650.

[21] W. Horz, W. Altenburger, *Nucl. Acids Res.* **9** (1981) 2643–2658.

[22] C. Dingwall, G. P. Lomonossof, R. A. Laskey, *Nucl. Acids Res.* **9** (1981) 2659–2673.

[23] J. T. Flick, J. C. Eissenberg, S. C. R. Elgin, *J. Mol. Biol.* **190** (1986) 619–633.

[24] M. Noll, R. D. Kornberg, *J. Mol. Biol.* **109** (1977) 393–404.