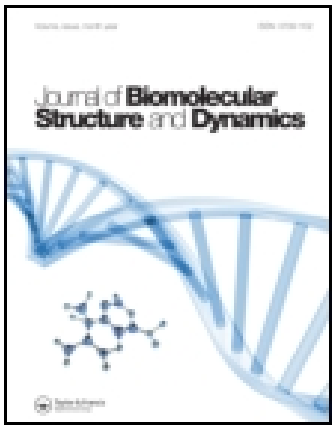

Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

Predicting Nucleosome Positions in Yeast: Using the Absolute Frequency

Zhiqian Zhang [a] , Yusen Zhang [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
Published online: 11 Jul 2012.

To cite this article: Zhiqian Zhang , Yusen Zhang & Ivan Gutman (2012) Predicting Nucleosome Positions in Yeast: Using the Absolute Frequency, Journal of Biomolecular Structure and Dynamics, 29:5, 1081-1088, DOI: 10.1080/073911012010525032

To link to this article: http://dx.doi.org/10.1080/073911012010525032

Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 29, Issue Number 5, (2012)

Predicting Nucleosome Positions in Yeast: Using the Absolute Frequency

http://www.jbsdonline.com

Zhiqian Zhang[1]
Yusen Zhang[1]*
Ivan Gutman[2]

[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

Abstract

Nucleosome is the basic structure of chromatin in eukaryotic cells, and they form the chromatin fiber interconnected by sections of linker DNA. Nucleosome positioning is of great significance for gene transcription regulation. In this paper, we consider the difference of absolute frequency of nucleotides between the nucleosome forming and nucleosome inhibiting sequences. Based on the 2-mer absolute frequency of nucleotides in genome, a new model is constructed to distinguish nucleosome DNA and linker DNA. When used to predict DNA potential for forming nucleosomes in *S. cerevisiae*, the model achieved a high accuracy of 96.05%. Thus, the model is very useful for predicting nucleosome positioning.

Key words: Nucleosome position; Nucleosome DNA; Linker DNA; 2-mer absolute frequency.

Introduction

Seventy-five to ninety percent eukaryotic genomic DNA is packaged in nucleosomes which are the basic repeating units. Each nucleosome contains approximately 165 bp genomic DNA, and the core nucleosome is about 147 bp genomic DNA which wrapped in 1.75 turns around a octamer of the histone proteins H2A, H2B, H3, and H4 (1). Two neighboring nucleosomes are joined by a stretch of around 20-80 bp free DNA, termed "linker DNA" (2). Nucleosome core DNA's precise location in genomic DNA is the nucleosome's positioning, which plays a very important role in many biological processes, including replication, transcription, DNA repair, *etc.* (3-12). How DNA sequence and deformation affect the positioning of nucleosomes has been the subject of extensive coverage recently (13-28) in this Journal.

Genomic DNA sequences demonstrate high variability in their binding affinity to the nucleosome core. A numerous attempts have been undertaken to find sequence-dependent signals on DNA determining the location and distribution of nucleosome and build the nucleosome positioning prediction models using various models. Several studies have provided extensive evidence indicating a sequence dependent positioning of nucleosomes along DNA (29). The ability of DNA to form nucleosomes depends, at least partially, on the underlying sequence (30). In previous works Satchwell detected a periodicity of 10 bp in chicken nucleosome sequence (31), Segal *et al.* defined this periodic signal as the nucleosome positioning code. Much work has been done to elucidate the nucleosome positioning signals that determine the preference of a particular region to bind to histones and form a nucleosome. The CA dinucleotide has been shown to be important for nucleosome positioning, and the decamer TATAAACGCC has a high affinity for histones (32, 33). With the high-throughput techniques such as chromatin

*Corresponding author:
Zhiqian Zhang
*E-mail: zhangys@sdu.edu.cn

immunoprecipitation (CHIP), coupled with microarrays (CHIP-chip) and CHIP coupled with sequencing techniques (CHIP-Seq), nucleosome positioning maps in genome have been obtained with high resolution for yeast. Segal *et al*. in 2006 (34) used a hidden Markov model for constructing a "nucleosome-DNA interaction model". Their model has a 50% predicting accuracy. In 2008, Yuan *et al.* proposed an N-score model to discriminate nucleosome and linker DNAs using wavelet energies as covariates in a logistic regression model (35). In the same year, a web-interface called "nuScore" was developed (36) for estimating the affinity of histone core to DNA and prediction of nucleosome positioning, based on the DNA deformation energy score. However, these methods are so complex that it will take a long time to test all chromosome sequences. Therefore, developing a more simple and accurate prediction method is purposeful.

In this paper, we point out a new method for nucleosome positioning prediction, based on the 2-mer absolute frequency of nucleotides in genome (37, 38). Our model for distinguishing nucleosome and linker DNAs in yeast genome has better performance comparing to the previous works. High predictive success rates are obtained by the new model, and in addition its computation is simpler than before.

Methods

Genomic DNA and Nucleosome Positioning Data

The genome sequences of Saccharomyces cerevisiae were downloaded from the *S. cerevisiae* Genome Database. The nucleosome positioning information was taken from published experimental data (39, 40). Positive dataset and negative dataset, which consist of the nucleosome forming and nucleosome inhibiting sequences, respectively, were constructed based on the work of Lee *et al.* (39). In the work, a nucleosome formation score was given to each of 1,206,684 DNA fragments occupying 81% of the *S. cerevisiae* genome based on microarray hybridization to nuclease digested genomic DNA. In order to compare with previous studies, a total of 9,900 fragments of 150 bp, having the highest hybridization scores, were selected as the nucleosome forming sequences to construct the positive set. Additional 9,900 fragments of 150 bp, having the lowest hybridization scores, were selected as nucleosome inhibiting sequences to construct the negative set. For more information, refer to (39-42).

2-mer Absolute Frequency

Many studies have shown that the frequency of the 2-mer nucleotide in DNA sequences has a great influence on the distribution of the nucleosomes, and the 2-mer nucleotide compositions of the nucleosome forming and of nucleosome inhibiting sequences are different. We analyze the nucleosome forming and of nucleosome inhibiting data that is taken from published experimental data (42).

We count the occurrences of dinucleotides in the positive and negative dataset. The average occurrences of dinucleotides in the two datasets are obtained from its occurrences by dividing the total number of the nucleosome forming sequences in the positive dataset and nucleosome inhibiting sequences in the negative dataset, respectively, which are shown in Table I. Observing Table I, we can find the 2-mer nucleotide compositions of the nucleosome forming and nucleosome inhibiting sequences are quit different. Furtherly, 1000 sequences were randomly selected from the nucleosome forming and nucleosome inhibiting data, respectively. By counting the numbers of 2-mer nucleotides, we can find some 2-mer nucleotides are different between nucleosome forming and the nucleosome inhibiting data, for example, the differences of AA and TA is shown in Figures 1 and 2.

Table I
The average occurrences of dinucleotides in two data sets.

Average occurrence	AA	AC	AG	AT	CA	CC	CG	CT
Nucleosome	12.188	9.395	9.754	9.616	11.831	7.613	4.375	10.204
Linker DNA	23.465	6.451	6.906	16.736	6.951	3.694	3.892	6.775
Average occurrence	GA	GC	GG	GT	TA	TC	TG	TT
Nucleosome	10.027	6.7014	7.131	9.123	6.916	10.310	11.721	12.095
Linker DNA	6.995	4.276	3.848	6.522	16.157	6.893	6.995	22.445

In this work, we consider the 2-mer absolute frequency (37). For a sequence S, a 2-bp sliding window was used to scan the DNA sequence from start to end in 1-bp steps. Then we can get the total numbers of sixteen 2-mer nucleotides N_x, where XY is one of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG and TT. The respective frequency of the 2-mer XY is

$$F_{XY} = N_{XY} / N \quad [1]$$

In the same way, we can define the frequency of a single nucleotide

$$F_X = N_X / N \quad [2]$$

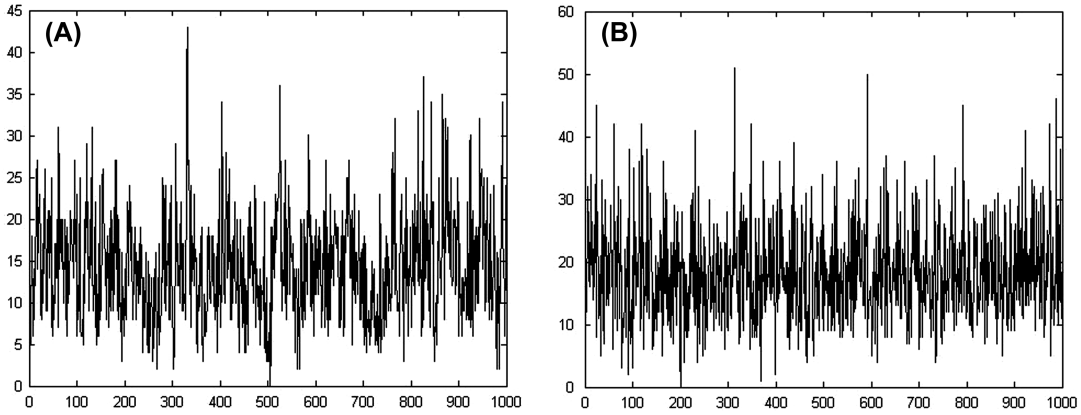

Figure 1: (**A**) The AA 2-mer nucleotides distribution of 1000 nucleosome forming data. (**B**) The AA 2-mer nucleotides distribution of 1000 nucleosome inhibiting data.

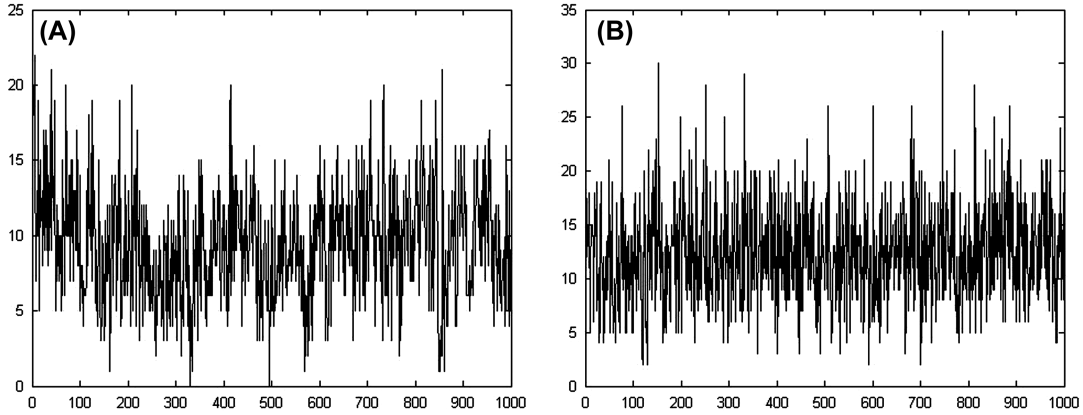

Figure 2: (**A**) The TA 2-mer nucleotide distribution of 1000 nucleosome forming data. (**B**) The TA 2-mer nucleotide distribution of 1000 nucleosome inhibiting data.

Where X is A, C, G and T, N is the total number of nucleotides of the sequences S. The 2-mer absolute frequency $P_S(XY)$ is defined as the ratio of the frequency of 2-mer XY to that of the first nucleotide X. That is

$$P_S(XY) = F_{XY} / F_X \quad [3]$$

So the 2-mer absolute frequency of sequences S, is defined by

$$x_S = [P_S(\mathrm{AA}), P_S(\mathrm{AC}), P_S(\mathrm{AG}), ..., P_S(\mathrm{TG}), P_S(\mathrm{TT})] \quad [4]$$

Thus, we get a sixteen-dimensional vector X_S, representing the DNA sequence.

Table II shows the 16 average 2-mer absolute frequencies of two data sets. We can observe that most of 16 average 2-mer absolute frequencies of nucleosome forming sequences in positive dataset are between 0.2 and 0.3 and that of nucleosome inhibiting sequences in negative dataset are more than 0.3 or less than 0.2. This same result happens when a fixed number of sequences are randomly selected from the positive and negative dataset, respectively.

The Algorithm Model

Using theories mentioned above, every sequence of the positive dataset and the negative dataset can be represented by corresponding sixteen-dimensional vector.

Next, we divide the positive dataset and the negative dataset into two parts: positive training set and positive test set, negative training set and negative test set, respectively. Positive test set and negative test set compose the test set. If a sequence is the i_th sequence of the positive training set, we can get a correspondence between the i_th sequence and an eighteen-dimensional vector

$$\mathrm{X}_S = [\mathrm{X}_{i1}, \mathrm{X}_{i2}, ..., \mathrm{X}_{i16}] \quad [5]$$

Where X_{ij} (j = 1, 2,…, 16) is the 2-mer absolute frequency of the i_th sequence. In this way, we can get a matrix X: $\mathrm{X} = \mathrm{X}_{ij}$ (i = 1, 2, …, n_1, j = 1, 2, …,16), where n_1 is the number of sequences in the positive training set. Set u_1 as the average vector of X. Then

$$\mathrm{u}_1 = [x'_1, x'_2, ..., x'_{16}]$$

and

$$x'_j = \sum_{i=1}^{n_1} x_{ij} / n_1 \quad [6]$$

Table II
The average 2-mer absolute frequency in two data sets.

Average absolute frequency	P (AA)	P (AC)	P (AG)	P (AT)	P (CA)	P (CC)	P (CG)	P (CT)
Nucleosome	0.2859	0.2270	0.2395	0.2399	0.3454	0.2111	0.1136	0.3028
Linker DNA	0.4151	0.1250	0.1310	0.3223	0.3368	0.1593	0.1755	0.3206
Average absolute frequency	P (GA)	P (GC)	P (GG)	P (GT)	P (TA)	P (TC)	P (TG)	P (TT)
Nucleosome	0.3034	0.2079	0.2027	0.2765	0.1731	0.2513	0.2853	0.2812
Linker DNA	0.3252	0.1919	0.1644	0.3108	0.3156	0.1330	0.1388	0.4064

Where, (i = 1, 2,…, n_1, j = 1, 2,…, 16). In the same way, we can get a matrix Y and the average vector u_2 from the negative training set:

$$Y = (y_{ij})\ (i = 1, 2,\ldots, n_2, j = 1, 2,\ldots,16)$$

$$u_2 = [y'_1, y'_2,..., y'_{16}]$$

Where n_2 is the number of sequences in the negative training set. Thirdly, for any sequence of the test set, we also can get a sixteen-dimensional vector T. In order to decide whether the sequence belongs to positive set or to negative set, a nonlinear discriminant function ξ is defined by the follow model:

$$\eta_1 = (T-u_1)\ (T-u_1)^T \quad [7]$$

$$\eta_2 = (T-u_2)\ (T-u_2)^T \quad [8]$$

$$\xi = (\eta_1-\eta_2)/2 \quad [9]$$

The parameter ξ gives the decision for the sequence. In this model, the threshold ξ_0 is decided by the training set. As $\xi > \xi_0$, the sequence S is classified into the positive group, whereas into the negative group as $\xi < \xi_0$.

Assessment of Prediction Performance

In order to evaluate the performance of a predictive algorithm, selecting a test method is an important issue. In the previous papers, the jackknife test and ROC curve were used normally. Both positive data set and negative data set were divided into five sets randomly. The first set, both positive and negative data set, was used as a test set, and the other four sets were retained as training sets. The model was evaluated by the test set. In order to test the robustness of the model, we select the 2nd, 3rd, 4th, and 5th set as test sets in turn, to evaluate the performance of the model and to use other retained four sets as training set, respectively. So, we prepared five training sets and five test sets and constructed five models.

The overall prediction accuracy (A) of the five models is defined as

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad [10]$$

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

$$S = \frac{TP}{TP + FN} \quad [11]$$

$$T = \frac{TP}{TP + FP} \quad [12]$$

and

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad [13]$$

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

Performance of the model was evaluated with 5-fold cross validation, and the quality of the classifier was shown in Table III. We also compare our method with IDQD methods (41) on the *S. cerevisiae* genome data set in jackknife test. Our model

Table III

Prediction results of the nucleosome core DNA and linker DNA in *S. cerevisiae* genome in jackknife test.

Test set	Accuracy (%)	Sensitivity (%)	Precision (%)	MCC (%)
Set 1	96.14	96.11	96.16	92.27
Set 2	96.06	96.06	96.06	92.12
Set 3	95.88	95.86	95.91	91.77
Set 4	96.14	96.11	96.16	92.27
Set 5	96.04	96.01	96.06	92.07
Average	96.05	96.03	96.07	92.1
IDQD	94.94	94.29	95.53	89.9

can distinguish the nucleosome core DNA sequence and linker DNA sequence in *S. cerevisiae* genome with high accuracy. The model achieved a good performance with average accuracy of 96.05% and MCC of 92%.

The ROC curve of the model in discriminating nucleosome core DNA and linker DNA sequences. 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 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. Our model's ROC cures is shown in Figure 3, and the AUC is shown in Table IV. Field *et al*. and Yuan *et al*. distinguished the nucleosome core DNA sequences and linker DNA sequences previously, the AUC were 0.975 and 0.955, respectively. The AUC of our model is 0.991, indicating higher accuracy of our model.

Discussion

Previous researches have shown that the form of nucleosome distribution has preference of DNA sequences; 2-mer nucleotide plays a very important role in it. That this is the case has been known for at least two decades. Travers *et al.* (43) showed in 1987, using statistical analysis of 177 DNA molecules from chicken erythrocytes, that certain 2-mer and 3-mer nucleotides within nucleosome-bound

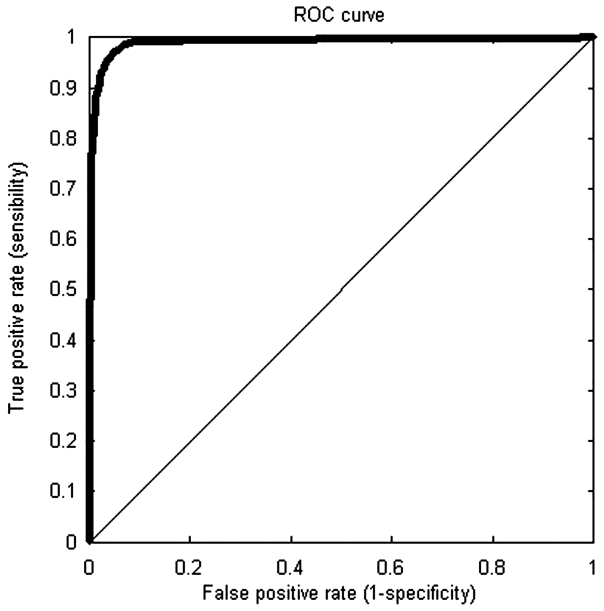

Figure 3: Relative Operating Characteristic curve (ROC).

Table IV
ROC curve analysis.

Test set	AUC	S.E.	95% C.I.	Comment
Set 1	0.99098	0.00151	0.98801-0.99394	Excellent test
Set 2	0.99082	0.00153	0.98783-0.99381	Excellent test
Set 3	0.99046	0.00156	0.98741-0.99351	Excellent test
Set 4	0.99074	0.00153	0.98774-0.99375	Excellent test
Set 5	0.99156	0.00146	0.98869-0.99442	Excellent test

sequences exhibit well-defined periodicities, based on the way specific bases influence the molecule's physical structure. Kunkel and Martinson discovered that poly (dA) -poly (dT) sequences are disfavored by nucleosomes and cannot appear in them at all above certain lengths (44). Later many models based on one of these signals (45-47) achieved limited success in prediction of nucleosome positions, but their results always get a lower accuracy. So developing new computational methods based on multiple factors is desirable. In this study, we consider the 2-mer absolute frequency of the nucleosome forming and nucleosome inhibiting data and highlights some difference between nucleosome DNA and linker DNA. Using the 2-mer absolute frequency to predict the core nucleosomes, we reach a more accurate result. However, 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 (48-51). If we add other factors to the vector, such as periodicity and curvature, results may be better. This is planned to be our next research in the future.

Acknowledgements

The author thanks the anonymous referees and editor for their corrections and valuable comments. This study was supported in part by the Shandong Natural Science Foundation (ZR2010AM020) and in part by the Serbian Ministry of Science and Education (Grant No. 174033).

References

1. C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond. *J Mol Biol 319*, 1097-1113 (2002)
2. K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, *Nature 389*, 251-260 (1997).
3. R. T. Simpson and D. W. Stafford. *Proc Natl Acad Sci USA 80*, 51-55 (1983).
4. B. Piña, U. Bruggemeier, and M. Beato. *Cell 60*, 719-731 (1990).
5. P. T. Lowary and J. Widom. *J Mol Biol 276*, 19-42 (1998).
6. A. Flaus, K. Luger, S. Tan, and T. J. Richmond. *Proc Natl Acad Sci USA 93*, 1370-1375 (1996).
7. A. Almer and W. Hörz. *Embo J 5*, 2689-2696 (1986).
8. H. Richard-Foy and G. L. Hager. *Embo J 6*, 2321-2328 (1987).
9. M. Shimizu, S. Y. Roth, C. Szent-Gyorgyi, and R. T. Simpson. *Embo J 10*, 3033-3041 (1991).
10. L. Verdone, G. Camilloni, E. Di Mauro, and M. Caserta. *Mol Cell Biol 16*, 1978-1988 (1996).
11. T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh. *Genome Res 18*, 1073-1083 (2008).
12. Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. *PLoS Comp Biol 4*, 1-25 (2008).
13. A. Travers, E. Hiriart, M. Churcher, M. Caserta, and E. Di Mauro, *J Biomol Struct Dyn 27*, 713-724 (2010).
14. F. Xu and W. K. Olson. *J Biomol Struct Dyn 27*, 725-739 (2010).
15. E. N. Trifonov. *J Biomol Struct Dyn 27*, 741-746 (2010).
16. P. De Santis, S. Morosetti, and A. Scipioni. *J Biomol Struct Dyn 27*, 747-764 (2010).
17. G. A. Babbitt, M. Y. Tolstorukov, and Y. Kim. *J Biomol Struct Dyn 27*, 765-780 (2010).
18. D. J. Clark. *J Biomol Struct Dyn 27*, 781-793 (2010).
19. S. M. Johnson. *J Biomol Struct Dyn 27*, 795-802 (2010).
20. G. Arya, A. Maitra, and S. A. Grigoryev. *J Biomol Struct Dyn 27*, 803-820 (2010).

21. F. Cui and V. B. Zhurkin. *J Biomol Struct Dyn 27*, 821-841 (2010).
22. D. Wang, N. B. Ulyanov, and V. B. Zhurkin. *J Biomol Struct Dyn 27*, 843-859 (2010).
23. S. M. West, R. Rohs, R. S. Mann, and B. Honing. *J Biomol Struct Dyn 27*, 861-866 (2010).
24. Y. V. Sereda and T. C. Bishop. *J Biomol Struct Dyn 27*, 867-887 (2010).
25. I. Gabdank, D. Barash, and E. N. Trifonov. *J Biomol Struct Dyn* 26, 403-411 (2009).
26. I. Gabdank, D. Barash, and E. N. Trifonov. *J Biomol Struct Dyn* 28, 107-121 (2010).
27. A. E. Rapoport, Z. M. Frenkel, and E. N. Trifonov. *J Biomol Struct Dyn 28*, 567-574 (2011).
28. A. Travers, E. Hiriart, M. Churcher, M. Caserta, and E. Di Mauro. *J Biomol Struct Dyn 27*, 713-724 (2010).
29. E. N. Trifonov. *Nucleic Acids Res 8*, 4041-4053 (1980).
30. H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng, *Genome Res 17*, 1170-1177 (2007).
31. S. C. Satchwell, H. R. Drew, and A. A. Travers, *J Mol Biol 191*, 659-675 (1986).
32. H. R Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson1, P. E. Nielsen, J. D. Kahn, D. M. Crothers, and M. Kubista, *J Biol Chem 267*, 807-817 (1997).
33. H. R. Widlund, P. N. Kuduvalli, M. Bengtsson, H. Cao, T. D. Tullius, and M. Kubista, *J Biol Chem 274*, 31847-31852 (1999).
34. E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. Moore, J. P. Wang, and J. Widom, *Nature 442*, 772-778 (2006).
35. G. C. Yuan and J. S. Liu, *PLoS Comput Biol 4*, e13 (2008).
36. M. Tolstorukov, V. Choudhary, W. Olson, V. Zhurkin, and P. Park, *Bioinformatics 28*, 1456-1458 (2008).
37. Y. S. Zhang and W. Chen, *J Biomol Struct Dyn 28*, 557- 565 (2011).
38. Y. S. Zhang and W. Chen, *MATCH Commun Math Comput Chem, 65*, 477-488 (2011).
39. W. Lee, D. Tillo, N. Bray, R. H. Morse, and R. W. Davis, *Nat Genet 39*, 1235-1244 (2007).
40. X. J. Zhao, Z. Y. Pei, J. Liu, S. Qin, and L. Cai, *Chromosome Res 18*, 777-785 (2010).
41. G. C.Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando, *Science 309*, 26-630 (2005).
42. X. F. Yi, Y. D.Cai, Z. S. He, W. R. Cui, and X. Y. Kong, *PLoS ONE 5(9)*, e12495 (2010).
43. A. A. Travers and A. Klug, *Phil Trans R Soc Lond 317*, 537-561 (1987).
44. G. R. Kunkel and H. G. Martinson, *Nucleic Acids Res 9*, 6869-6888 (1981).
45. T. Schalch, S. Duda, D. F. SargentF, and T. J. Richmond, *Nature 436*,138-141 (1997).
46. H. Cao, H. R. Widlund, T. Simonsson, and M. Kubista, *J Mol Biol 281*, 253-260 (1998).
47. Y. Field, Y. Fondufe-Mittendorf, and I. K. Moore, *Nat Genet 41*, 438-445 (2009).
48. W. Hörz and W. Altenburger. *Nucl Acids Res 9*, 2643-2658 (1981).
49. C. Dingwall, G. P. Lomonossof, and R. A. Laskey. *Nucl Acids Res 9*, 2659-2673 (1981).
50. J. T. Flick, J. C. Eissenberg, and S. C. R. Elgin. *J Mol Biol 190*, 619-633 (1986).
51. M. Noll and R. D Kornberg. *J Mol Biol 109*, 393-404 (1977).

Date Received: July 8, 2010

Communicated by the Editor Ramaswamy H. Sarma