Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Hypothesis**

Prediction of nucleosome positions in the yeast genome based on matched mirror position filtering

Qinqin Wu[1,2,*], Jiajun Wang[2] and Hong Yan[1,3]

[1]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong; [2]School of Electronics and Information Engineering, Soochow University, Suzhou, China; [3]School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia; Qinqin Wu* - Email: suzhouqinqin@gmail.com; *Corresponding author

Received March 11, 2009; revised May 01, 2009; accepted June 22, 2009; published August 06, 2009

Abstract:
Nucleosome positioning can affect the accessibility of the underlying DNA to the nuclear environment and as such plays an essential role in the regulation of cellular processes. Specific patterns have been found in the underlying DNA sequences of the nucleosome, and one of the most important patterns includes dinucleotides distributed every 10 to 11 base pairs. Based on this property, we propose to match each dinucleotide in the sequence against its mirror occurrences for 10 to 11 base pairs on both left-hand and right-hand sides. A large number of matches in a local region will then signify the existence of a nucleosome. In this paper, we propose the matched mirror position filters for efficient matching of periodic dinucleotide patterns and computationally predict the nucleosome positions. Experimental results on the Saccharomyces cerevisiae (yeast) genome show that the proposed algorithm can predict nucleosome positions effectively. More than 50% of our predicted nucleosomes are within 35 base pairs of those detected by biological experiments.

Keywords: mirror position filter; nucleosome positioning; Saccharomyces cerevisiae; yeast genome; DNA

Background:
The nucleosome, which is the primary unit of a eukaryotic chromatin, contains about 147 base pairs (bp) of DNA which are sharply bent and tightly wrapped around a histone octamer. This sharp bending occurs at every DNA helical repeat (~10 bp) where the major groove of the DNA faces inwards towards the histone octamer, and again ~5 bp away, with opposite direction, the major groove faces outward **[1].** Along the chromatin, the neighboring nucleosomes are separated from each other by 10 to 50-bp long stretches of unwrapped linker DNA [Figure 1], thus, 75–90% of genomic DNA is wrapped in nucleosomes and as such affects sequence accessibility **[2].** The positioning of nucleosomes along the chromatin is known to play an important role in the regulation of gene expression in eukaryotic cells.

Extensive research has been carried out on nucleosome positioning recently. Some of the research tries to explore the mechanic properties of nucleosomes, while the others develop mathematical models for the prediction of the nucleosome positions **[1, 3-5].** In their study of nucleosome placement in chromatin, Vincent Miele proposed that the physical properties of DNA may determine the nucleosome occupancy from yeast to fly **[6].** Besides, I.P. Ioshikhes predicted some nucleosomes on the basis of comparative genomics while M. Yassour located the nucleosomes by analyzing microarray data **[7, 8].** Despite the fact that genome-wide maps of nucleosome locations have been generated **[2, 6, 9, 10],** the problem of accurately predicting the nucleosome positions computationally at high resolution remains unresolved.

Statistical analysis suggests that the periodicities of the underlying DNA sequence might help solve this problem **[2].** Indeed, there is evidence that the distinctive sequence motifs recurring periodically at the DNA helical repeat facilitate the sharp bending of DNA around the nucleosome and hence favor nucleosome formation. These motifs include ~10-bp periodic AA/TT/TA dinucleotides that oscillate in phase with each other and out of phase with ~10-bp periodic GC dinucleotides **[2, 4, 11].** Based on this property, we propose a computational method using the so-called matched mirror position filter (MMPF) for the prediction of nucleosome positions. An advantage of our method is that it does not require training data and thus expensive wet-lab experiments are not needed. The computational experiment results demonstrate that our approach can detect the positions of the nucleosome effectively. On average, more than 50% of our predicted stable nucleosomes are within 35bp of those detected by biological experiments and reported in literature.

Methodology:
Nucleotide base set:
Let us define the set B1 = {A, C, G, T}, which contains the four nucleotide bases. Then consequently there should be 4×4 =16 dinucleotides. S. C. Satchwell has validated that in the 16 possible dinucleotides only 10 of them are unique. He explained that it is because some of them are related to the two fold axis that passes between the two strands of the double helix, and the reverse complementary dinucleotides are considered to be equivalent **[11].** Considering this factor, we define the unique dinucleotides as set B2 = {AA/TT, AT, AC/GT, AG/CT, TA, TC/GA, TG/CA, CC/GG, CG, GC}.

Dirac delta function
Assume that a DNA sequence is represented by a discrete function $x(n)$, with $x(n) \in B_1$ and $n = 0, 1, 2, \ldots, N-1$. For each dinucleotide $b \in B_2$, a delta function is defined as in **Equation 1 (see supplementary material)**. The notation of this delta function has been used to represent the positions of nucleotide bases **[12, 13]**. Here we generalize the notation to dinucleotides. For example, if $x(n) =$

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)

Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Hypothesis**

{GCAGGTACAGTCGAG}, then $x_{AG}(n) = \delta(n-2) + \delta(n-8) + \delta(n-13)$.

Matched mirror position filter

Ideally, to produce a periodicity of ~10.5 bp, an impulse $\delta(n-n_k)$ at each position n_k should be like a double-side mirror. That is, it should reflect an impulse $\delta(n-n_{kl})$ on the left-hand side 10 to 11 bp away and an impulse $\delta(n-n_{k+r})$ on the right-hand side 10 to 11 bp away (**Figure 2**). As defined in **Equations (2)** and **(3)** (**see supplementary material**) respectively, $d_L(b, n_k)$ is the distance of $\delta(n-n_k)$ to the impulse closest to the position n_k 10.5, and $d_R(b, n_k)$ is the distance of $\delta(n-n_k)$ to the impulse closest to the position n_k 10.5. Then, ideally we should have $d_L(b, n_k) = 10$ or 11 and $d_R(b, n_k) = 10$ or 11 and there should be many of these impulses in a nucleosome to produce the periodicity of ~10.5 bp and no such impulse in a linker.

In practice, $d_L(b, n_k)$ and $d_R(b, n_k)$ may deviate from the ideal values. To take this deviation into account, we use a matching function $f(d)$ to measure the contributions of $\delta(n-n_k)$, $\delta(n-n_{kl})$ and $\delta(n-n_{kr})$ to the periodicity. The matching function should be large for d close to 10.5 and decrease as d moves away from this optimal value. Several choices of $f(d)$ are shown in **Figure 3**. Given the matching function, the contributions of the three impulses to the periodicity can be described as **Equation 4 (see supplementary material)**.

To detect the presence of nucleosomes, we move a window with the size of $2W_s + 1$ along the DNA sequence and accumulate the contributions from all dinucleotides within the window. Consequently, the nucleosome matching score function can be defined as **Equation 5 (supplementary material)**. In fact, our proposed method involves flexible matching of mirrored dinucleotide positions. To some extent, it is similar to the matched filter used in radar systems for detecting echo signals. Thus, our system is named the matched mirror position filter (MMPF) with $S(n)$ being the output.

Threshold discriminant

Following the ideas discussed above, the nucleosomes can be predicted by comparing $S(n)$ as defined in **Equation 5 (see supplementary material)** with a threshold distinguishing the nucleosome and the non-nucleosome. The threshold here is determined empirically by performing a lot of experiments on a number of nucleotide sequences. After a lot of experimental trials, we found that $T = 1.2(2W_s + 1)$ is the optimal threshold. In the implementation of our proposed method, the window length is selected as $2W_s + 1 = 147$ according to the fact that the lengths of nucleosomes in these sequences studied are mostly 147 bp **[2, 6, 9, 10]**. A window size in the range of 100 to 200 gives similar results, but the algorithm achieves the best result in terms of the false positives in the sequences studied with the window size of 147. For this window size, the threshold for the nucleosome matching score is $1.2(2W_s + 1) \approx 180$.

Discussion:

In order to verify the performance of the MMPF, we have performed nucleosome position prediction experiments on the *Saccharomyces cerevisiae* (yeast) genome, which is downloaded from the database of National Center for Biotechnology Information (NCBI) **[14]**. The genome contains 16 chromosomes with lengths from 230k to 1532k bp. We choose the yeast genome since it has been extensively studied and there are high-resolution nucleosome mapping results reported in literature [7]. A criterion similar to that in **[2]** is used here to compute the prediction accuracy. That is, we consider a prediction correct if the nucleosome center position determined by the MMPF is within 35 bp of that reported in the high-resolution result of [7]. The prediction accuracy is then computed as the ratio of correctly predicted nucleosomes over the total number of nucleosomes predicted by our algorithm.

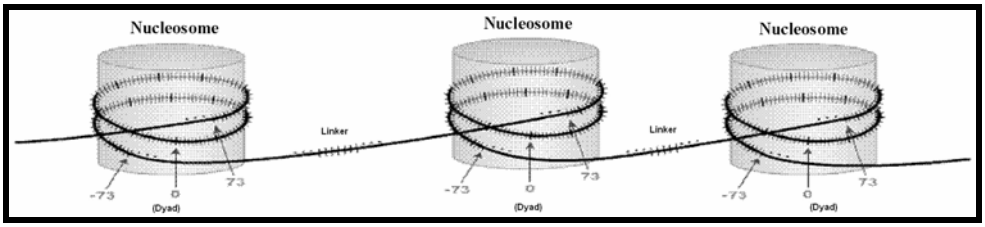

Figure 1: Examples of nucleosomes, each containing 147 bp of nucleotides, connected by linkers of different lengths.

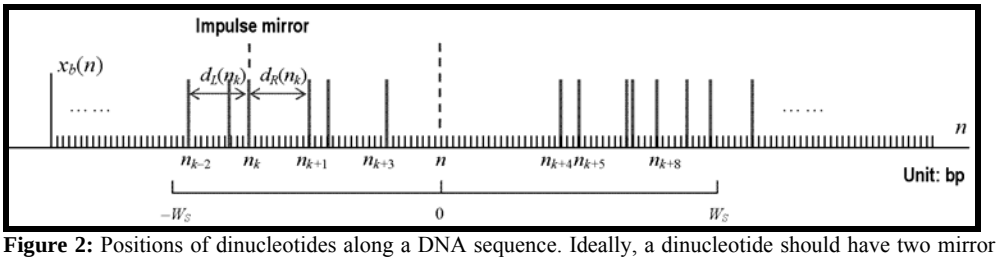

Figure 2: Positions of dinucleotides along a DNA sequence. Ideally, a dinucleotide should have two mirror images, on the left-hand and right-hand sides respectively, 10 to 11 bp away to produce the periodicity of ~10.5 bp in a nucleosome.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)

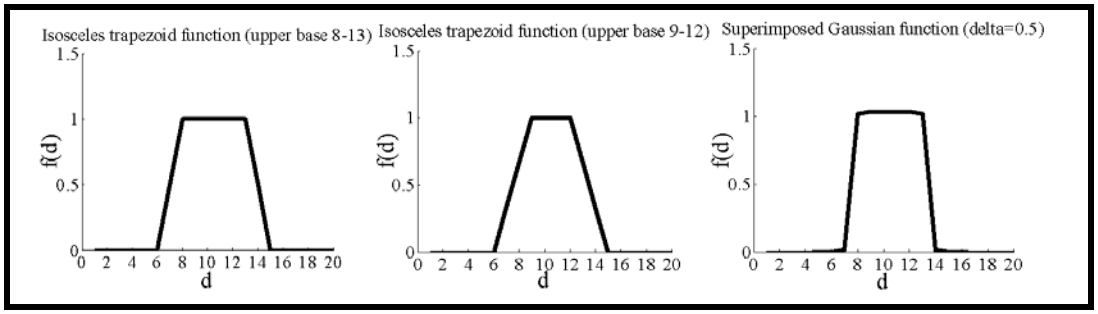

Figure 3: Possible choices of the matching function *f*(*d*) with *d* representing the distance from the current position. Diagrams in (a) and (b) show trapezoids with different base lengths. Diagram (c) shows a matching function consisting of the summation of six Gaussian functions with standard deviation equal to 0.5 and means at 8, 9, 10, 11, 12 and 13 bp.

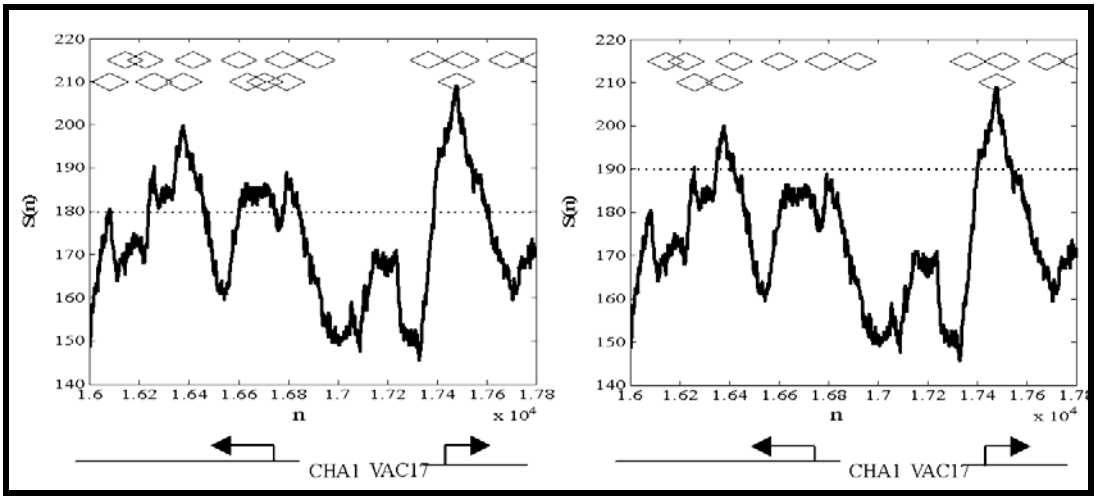

Figure 4: Genome-wide prediction of nucleosome positions in yeast chromosome III using the MMPF with *n* denoting the position in the chromosome and *S(n)* denoting the matching score. (a) The distribution of the nucleosomes near genes *CHA1* and *VAC17*. The upper green diamonds show the nucleosomes determined from biological experiments **[7]** and the lower red diamonds show the nucleosomes detected using our computational method. The dashed line represents the threshold at 180 for the nucleosome matching score. (b) The threshold is changed to 190.

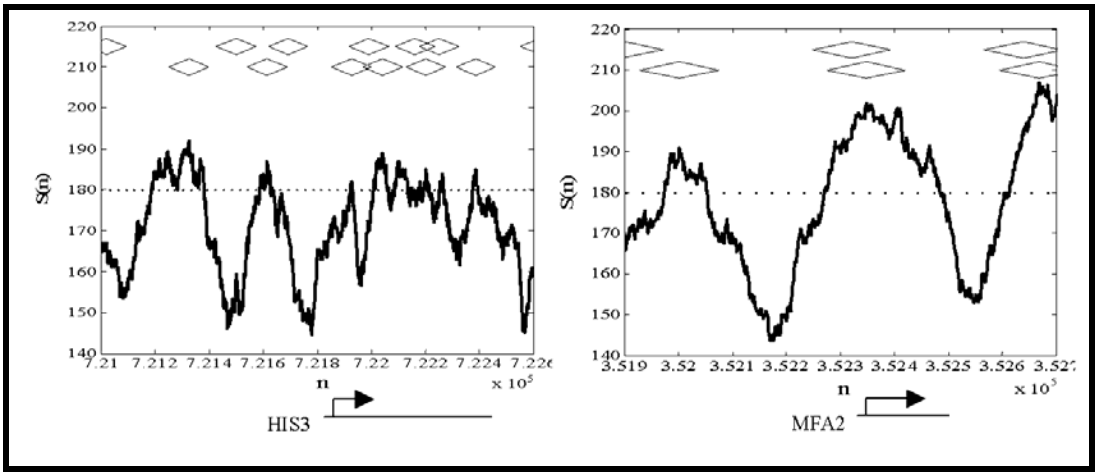

Figure 5: Genome-wide prediction of nucleosome positions in chromosome XV and chromosome XIV using the MMPF with *n* denoting the position in the chromosome and *S(n)* denoting the matching score: (a) near gene *HIS3* and (b) gene *MFA2*.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)

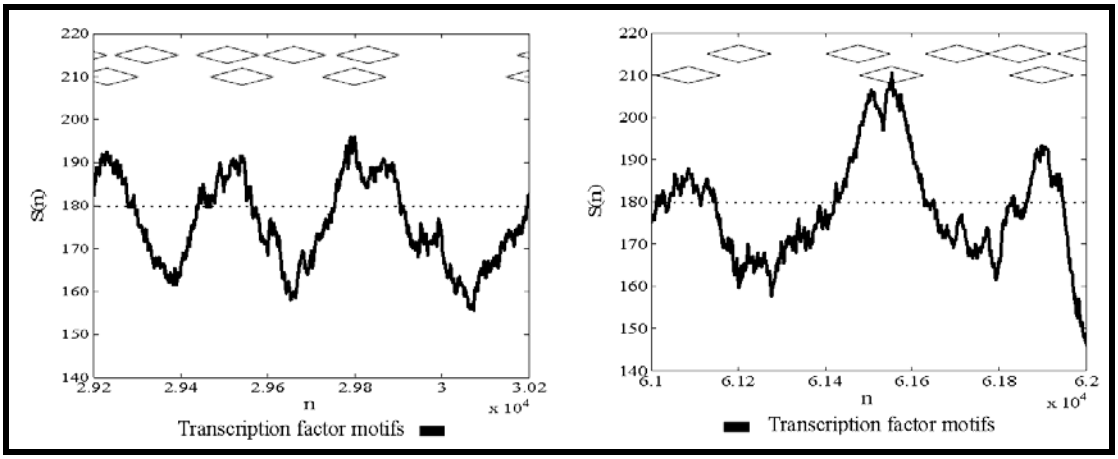

Figure 6: Genome-wide prediction of nucleosome positions near the transcription factor motifs **[15]** in chromosome XVI and chromosome III using the MMPF with *n* denoting the position in the chromosome and *S(n)* denoting the matching score.

The prediction results achieved by using different matching functions in the MMPF are summarized in **Table 1 (Supplementary material)**. From this table, it can be seen that with the isosceles trapezoid function (upper base 9-12) the least nucleosomes can be predicted and the accuracy is the worst, while with superimposed Gaussian function (delta = 0.5) the accuracy achieves the best occasionally but its predicted nucleosomes are still less than using the isosceles trapezoid function (upper base 8-13). In consideration of both the number of correctly predicted nucleosomes and the prediction accuracy, we choose the isosceles trapezoid function (upper base 8-13) as the optimal matching function *f(d)* to predict nucleosome positions for further analysis.

For a better understanding of the correlation between the predicted nucleosomes and the regulatory function sites, such as promoters and the transcription factor binding sites, a comparison between the nucleosomes predicted by our method and those obtained from biological experiments and published in literature is shown in **Figures 4, 5** and **6**. Besides, the effect of different thresholds on the accuracy is also discussed. From **Figure 4a**, it can be seen that in the segment near the genes *CHA1* and *VAC17* of chromosome III, ten nucleosomes are reported in literature. Seven stable ones are predicted with our proposed method, and four of them coincide well with those determined in biological experiments. In **Figure 4b**, when the threshold is set to be 190, although the number of false positives decreases, the accuracy of the whole genomic-scale prediction degrades to ~45%.

Figure 5 shows the comparison between our predicted stable nucleosomes and those reported in literature near the genes *HIS3* and *MFA2* in chromosomes XV and XIV, respectively. A similar comparison of the nucleosomes near some transcription factor motifs [**15**] (in chromosomes XVI and III) is shown in **Figure 6**. The results shown in **Figures 4, 5** and **6** suggest that the potential nucleosomes can be effectively predicted by the MMPF. Although in some segments the rate of correctly predicting nucleosomes is low (**Figure 6b**), this algorithm performs well on a genomic scale. Our analysis also indicates that nucleosomes may play a role in the regulation of the DNA sequences. From **Figures 4a** and **5**, it can be seen that there may be correlations existing between the positions of nucleosomes and genes. It can also be seen from these figures that the nucleosomes have a strong affinity to the genes in some genomic locations and exhibit lower occupancy in other positions such as the promoters, usually found ~100 to 500 bp upstream of the start codon in the gene or in the intergenic regions. Besides, our findings in **Figure 6** support the hypothesis that the functional transcription factor binding sites may be predominantly nucleosome-free **[15]**.

Conclusion:

In this paper, we have presented a computational method based on the MMPF for nucleosome position prediction. This technique, while offering a level of accuracy comparable to existing ones, has some distinct advantages. Being based on the periodicity of nucleosomes and a flexible pattern matching scheme, it is independent of a training set that must be obtained through biological experiments. The MMPF can provide a useful tool to study the eukaryotic genome chromatin structure, protein-DNA interactions, and transcriptional regulations. Our future work aims to improve the prediction accuracy for large genomic sequences based on more sophisticated mathematical models, such as probabilistic relaxation labeling [**16**].

Acknowledgement:

We thank Dr. Hongya Zhao for helpful suggestions. This work is supported by the Hong Kong Research Grant Council (Project CityU 123408), the National Natural Science Foundation of China (Project 60871086) and the Natural Science Foundation of Jiangsu Province, China (Project BK2008159).

References:

[1] T. J. Richmond & C. A. Davey, *Nature,* (2003)423: 145

[2] E. Segal, *et al., Nature,* (2006) 442: 772

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)

[3] K Chen et al., Nucleic Acids Res. (2008) 36:6228.
[4] M. Bina, *J. Mol. Biol.,* (1994) 235: 198
[5] H. E. Peckham, *et al.*, *Genome Res.,* (2007)**17**: 1170
[6] V. Miele, *et al., Nucleic Acids Research,* (2008) 36: 3746
[7] I. P. Ioshikhes, *et al., Nat. Genet.,* (2006) 38: 1104
[8] M. Yassour, *et al., Bioinformatics,* (2008) 24: 139
[9] G. C. Yuan, *et al., Science,* (2005) 309: 626
[10] D. K. Pokholok, *Cell,* (2005) 122: 517
[11] S. C. Satchwell, *et al., J. Mol. Biol.,* (1986) 191: 659
[12] R. F. Voss, *Phys. Rev. Lett.,* (1992) 68: 3805
[13] D. Anastassiou, *IEEE Signal Proc. Mag.,* (2001) 18: 8
[14] http://www.ncbi.nlm.nih.gov
[15] C. T. Harbison, *et al.*, *Nature,* (2004) 431: 99
[16] H. Zhao & H. Yan, *Proc. Int'l Multiconf. Engineers & Computer Scientists*, (2009) 245

Edited by P. Kangueane
Citation: Wu *et al***,** Bioinformation 3(10): 454-459 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)

Bioinformation by Biomedical Informatics Publishing Group *open access*
www.bioinformation.net **Hypothesis**

Supplementary material

Table 1: Nucleosome prediction results for the yeast genome

		Isosceles trapezoid function (upper base 8-13)			Isosceles trapezoid function (upper base 9-12)			Superimposed Gaussian function (delta=0.5)		
C	**L**	**NP**	**NCP**	**AC**	**NP**	**NCP**	**AC**	**NP**	**NCP**	**AC**
I	230	946	489	51.69%	269	107	39.78%	549	276	50.27%
II	813	3328	1632	49.04%	967	396	40.95%	1881	978	51.99%
III	317	1302	644	49.46%	419	157	37.47%	799	386	48.31%
IV	1532	6372	2804	44.01%	1842	665	36.10%	3671	1613	43.94%
V	577	2347	1143	48.70%	689	276	40.06%	1371	656	47.85%
VI	270	1120	566	50.54%	312	116	37.18%	643	327	50.86%
VII	1090	4506	2113	46.89%	1354	508	37.52%	2676	1183	44.21%
VIII	563	2357	1200	50.91%	679	292	43.00%	1325	677	51.09%
IX	440	1751	915	52.26%	553	223	40.33%	1026	545	53.12%
X	746	3102	1593	51.35%	892	392	43.95%	1815	964	53.11%
XI	666	2760	1271	46.05%	826	317	38.38%	1615	710	43.96%
XII	1078	4435	1959	44.17%	1285	457	35.56%	2526	1121	44.38%
XIII	924	3714	1668	44.91%	1122	376	33.51%	2205	938	42.54%
XIV	784	3230	1559	48.27%	921	361	39.20%	1799	853	47.42%
XV	1091	4438	2031	45.76%	1333	501	37.58%	2495	1139	45.65%
XVI	948	3873	1821	47.02%	1129	422	37.38%	2197	1027	46.75%

C = chromosome; L = chromosome length; NP = Number of predicted nucleosomes; NCP = Number of correctly predicted nucleo-somes; AC = Accuracy

Equation 1:

$$x_b(n) = \ldots+\delta(n - n_{k1})+ \delta(n - n_k)+ \delta(n - n_{k1})+\ldots \quad (1)$$

where $\delta(n - n_k)$ is an impulse represented by the Dirac delta function indicating the occurrence of b at position n_k.

Equation 2:

$$d_L(b,n_k) = n_k - n_{k-l}, \text{ where } l = \underset{l>0}{\operatorname{argmin}}\{|n_k - n_{k-l} - 10.5|\}. \quad (2)$$

Equation 3:

$$d_R(b,n_k) = n_{k+r} - n_k, \text{ where } r = \underset{r>0}{\operatorname{argmin}}\{|n_{k+r} - n_k - 10.5|\}. \quad (3)$$

Equation 4:

$$s(b,n_k) = f(d_L(b,n_k)) + f(d_R(b,n_k)). \quad (4)$$

Equation 5:

$$S(n) = \sum_{b \in B} \sum_{n-W_s \le n_k \le n+W_s} s(b, n_k). \quad (5)$$

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(10): 454-459 (2009)