A multi-layer method to study genome-scale positions of nucleosomes

Vito Di Gesù a, Giosuè Lo Bosco a,*, Luca Pinello a, Guo-Cheng Yuan b, c, Davide F.V. Corona d, e

a Dipartimento di Matematica ed Applicazioni, Via Archirafi 34, 90123 Palermo, Italy
b Department of Biostatistics, Harvard School of Public Health, USA
c Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, USA
d Istituto Telethon Dultecco, c/o Università di Palermo, Italy
e Dipartimento di Scienze Biochimiche, Università di Palermo, Italy

ARTICLE INFO

Article history:
Received 11 June 2008
Accepted 29 September 2008
Available online 21 November 2008

Keywords:
Nucleosome positioning
Hidden Markov Model
Classification
Multi-layer method

Abstract

The basic unit of eukaryotic chromatin is the nucleosome, consisting of about 150 bp of DNA wrapped around a protein core made of histone proteins. Nucleosomes position is modulated in vivo to regulate fundamental nuclear processes. To measure nucleosome positions on a genomic scale both theoretical and experimental approaches have been recently reported. We have developed a new method, Multi-Layer Model (MLM), for the analysis of nucleosome position data obtained with microarray-based approach. The MLM is a feature extraction method in which the input data is processed by a classifier to distinguish between several kinds of patterns. We applied our method to simulated-synthetic and experimental nucleosome position data and found that besides a high nucleosome recognition and a strong agreement with standard statistical methods, the MLM can identify distinct classes of nucleosomes, making it an important tool for the genome wide analysis of nucleosome position and function. In conclusion, the MLM allows a better representation of nucleosome position data and a significant reduction in computational time.

© 2008 Elsevier Inc. All rights reserved.

Introduction

Nucleosomes in eukaryotes wraps 150 bp DNA or about 1.7 turns and their positioning plays an important role in gene regulation [1]. While this packaging allows the cell to organize a large and complex genome in the nucleus, it can also block the access of transcription factors and other proteins to DNA [2]. For example, under normal conditions the Pho5 promoter in yeast is occupied by well-positioned nucleosomes, preventing the transcription factor Pho4 from binding to its target binding site. When induced by phosphate starvation, nucleosomes are depleted from the promoter region so that Pho4 can bind to its target DNA binding sequence thus activating the Pho5 gene transcription [3]. However, nucleosome binding can sometimes enhance transcription by bringing distant DNA regulatory elements together [4]. Genome-wide studies have found that transcription activity is inversely proportional to nucleosome depletion in promoter regions in general [5–7]. With the help of tiling arrays at 20 bp resolution, Yuan et al. [8] have looked at nucleosome occupancy relative to gene regulatory regions on 4% of the yeast genome by using an Hidden Markov Model approach (HMM). The used microarray-based method allows the identification of nucleosomal and linker DNA sequences on the basis of susceptibility of linker DNA to micrococcal nuclease. This method allows the representation of microarray data as a signal of green/red ratio values showing nucleosomes as peaks of about 150 bp long, surrounded by lower ratio values corresponding to linker regions. Consistent with previous studies, Yuan et al. found that 87% of the transcription factor binding sites [9] are free of nucleosome binding. A substantial improvement over this work has been recently done by Lee et al. [10] where the genome-wide nucleosome positions in yeast have been mapped at 4 bp resolution. A similar approach has also been used to look at differences in nucleosome spacing occurring in the absence of a chromatin remodeler [11]. A number of other groups have developed analysis methods to detect nucleosomes as well as transcription factor binding sites [12–19]. Compared to transcription factors, it is more challenging to detect nucleosome positions since the majority of a eukaryotic genome is wrapped into nucleosomes. Another difficulty is that the raw data may contain complex trends that are unrelated to nucleosome binding [8]. An intuitive method to deconvolve data trend is to define a peak-to-trough difference measure and to detect its local maxima. However, Yuan et al. [8] have found that although this method can detect local peaks, it suffers from amplifying observation noise. A similar approach has been adapted in [20] to map nucleosome positions in human. Although an intrinsic DNA code for nucleosome positioning has been recently reported [21], a significant technological development in genome-wide location of nucleosomes has been made using “deep sequencing” approaches [22–25], which differs from microarray-based approach in that the isolated DNA of interest is mapped to genome via direct DNA sequencing, instead of microarray hybridization. For this new technology, the input data correspond to peaks of DNA fragment counts instead of high hybridization ratio. However, the
task of peak detection remains a key problem for the statistical analysis of the input data. Unlike microarray-based approaches, where data collection is constrained to a regular grid, “deep sequencing” data are intrinsically base-pair resolution and therefore less statistically stable. One solution to this problem is to first map the data onto a regular grid by binning. However, more sophisticated methods need to be developed to balance the resolution vs variance dilemma. The analysis of stochastic signals aims to both extract significant patterns from noisy background and to study their spatial relations (periodicity, long term variation, burst, etc.). The problem becomes more complex whenever the noise background is structured and unknown. Examples of such kind of data correspond to protein-sequences in the study of gene expression [8]. The analysis carried out in both cases has been based on probabilistic networks [27] (for example, Hidden Markov Models [28], Bayesian networks). Methods based on probabilistic networks are suitable for the analysis of such kind of signal data; however, they suffer of high computational complexity and results can be biased from locality that depends on the memory steps they use [8,26].

We developed a new method, Multi-Layer Model (MLM), strongly related to the class of approaches successfully used in the analysis of very noisy data [29]. Using several views of the input data-set the MLM allows a better pattern shape characterization of the input data and a significant reduction in computational time over the Hidden Markov Model (HMM). We tested the MLM to both synthetic and microarray-based nucleosome positioning data [8] and found that our method can identify several classes of positioned nucleosomes. Distinct nucleosome positions can underline important regulatory roles, highlighting the impact our method can have on genome-wide nucleosome phasing studies in higher eukaryotes.

**Note:** The MLM package including a short documentation, the software implemented in MatLab 6.5, and samples of input data can be downloaded from the webpage: http://www.math.unipa.it/pinello/mlm.

### Materials and methods

The MLM analysis is performed on both emulated and real signals; in both cases such signals comes from a microarray where each spot represents a probe $i$ of $r$ base pairs that overlaps every $o$ base pairs with probe $i+1$. In particular, the chromosome is spanned by moving a window (probe) $i$ of width $r$ base pairs from left to right, measuring both the percentage of mononucleosomal DNA $G$, (green channel) and whole genomic DNA $R_i$ (red channel) within such window, respecting also that two consecutive windows (probes) have an overlap of $o$ base pairs. The resulting signal $V(i)$ for each probe $i$ is the logarithmic ratio of the green channel $G_i$ to red channel $R_i$. Intuitively, nucleosomes presence is related to peaks of $V$ which correspond to higher logarithmic ratio values, while lower ratio values show nucleosome free regions called linker regions (see Fig. 1a). Note that, since the overlapping zone of the tiling microarray is $o$ bp, nucleosomes closer than this value will be not classified as well positioned but fused or delocalized (see Nucleosome classification section for more details and Fig. 1b as an example of nucleosomes region classes). The real signal that has been analyzed comes from the Saccharomyces cerevisiae chromosome and information about the used microarray labeling and hybridization protocols can be found in [8]. The MLM is based on the generation of several sub-samples of the input signal and in particular several thresholds, chosen by respecting cut-set optimal conditions, are applied to the input data. MLM is a general pattern detection method and it can be adapted to discover patterns on one-dimensional signals.

**Nucleosome identification**

The input microarray data, $S$, are organized in $T$ contiguous fragments $S_1,\ldots,S_T$ which represents DNA sub-sequences. In the following, a detailed description of the MLM processing steps is provided.

**MLM preprocessing step**

A preprocessing is necessary in order to reduce the effect of the signal noise. Each fragment $S_i$, $1 \leq i \leq T$ of the input signal, $S$, is smoothed by a convolution operator that perform the weighted average of three consecutive signal values, where the weights are provided by a kernel window $w = \frac{1}{3}, \frac{1}{2}, \frac{1}{3}$ [30].

**MLM model construction step**

Since we know that well positioned nucleosomes are shown as peaks of a bell shaped curve, in order to locate the position of a nucleosome, all local maxima of the input signal are automatically extracted from the convolved signal $X$ of $S$. Then a subset of maxima
The selection process extracts the significant patterns are shown. This further selection criterion is related to the height of the shaped bell fragment, in fact a small value of m could represent noise rather than nucleosomes. The value m is said the minimum number of permanences; in Parameter selection by calibration section a calibration procedure to estimate the best value of m is described.

MLM feature extraction step

Each pattern \( P_i \in P^m \) is identified by \( I^0_{fl}, I^+_{fl}, \ldots, I^{-l}_{fl} \), with \( l \geq m \). Straightforwardly, the feature vector of \( P_i \) is a \( 2 \times l \) matrix where each column represents the lower and upper limits of each interval from the lower threshold \( j \) to the upper threshold \( j+l \). The representation in this multi-dimensional feature space is used to characterize different types of patterns.

MLM dissimilarity function

A dissimilarity function between patterns is defined in order to characterize their shape:

\[
\delta(P_1, P_2) = (1-\alpha)(A_r-A_l) + \alpha \sum_{i=1}^{2} (a_i^c-a_i^l)
\]

where, \( A_r \) and \( A_l \) are the surfaces of the two polygons bounded by the set of vertices \( V = \cup_{i=1}^{2} \{ (b_i^c, c_i^c), (b_i^l, c_i^l) \} \), \( a_i^c \) and \( a_i^l \) are the x-coordinates of the (convolved) signal, and \( \alpha \) is a user parameter ranging in the interval \([0,1] \) to set the weight of the two dissimilarity components.

The first component of this dissimilarity allow us to consider patterns of close dimensions, while the second component has been introduced to introduce shape information in fact it can be considered a correlation measure of the two bounding polygons. This dissimilarity can be used by a general classifier in order to distinguish the kind of pattern. An example of input signal and the extracted interesting patterns is given in Fig. 1.

Nucleosome classification

MLM able to classify four kind of patterns: linkers, well positioned, delocalized and fused nucleosomes. (see Fig. 3(a)).

In the following, the classification rules which allow us to automatically discriminate such kind of patterns are stated. The classification was conducted in two steps, in the first step the linker patterns, the expected well positioned patterns and expected delocalized patterns are found. Afterwards, the ranges of the regions representing the expected well positioned and delocalized nucleosomal patterns

\[
\begin{align*}
&\text{P} = \{ I^0, I^+_{fl}, \ldots, I^{-l}_{fl} | \forall \text{intervals } I^0, I^+_{fl}, \ldots, I^{-l}_{fl} \} \\
&\text{P} = \{ I^0_{fl}, I^+_{fl}, \ldots, I^{-l}_{fl} | \forall \text{intervals } I^0_{fl}, I^+_{fl}, \ldots, I^{-l}_{fl} \}
\end{align*}
\]

where, \( j \) defines the threshold, \( j_p \) of the widest interval of the pattern. From the previous definition it follows that \( P_i \) is build by adding an interval \( I^0_{fl} \) only if it is the unique in \( R_{k+1} \) that is included in \( I^0_{fl} \). Note that, this criterion is inspired by the consideration that a nucleosome is identified by bell shaped fragment of the signal, and the intersection of such fragment with horizontal threshold lines results on a sequence of nested intervals. In Fig. 2 two examples of shapes with the relative patterns are shown.

MLM pattern selection step

In this step the interesting patterns \( P^m \) are selected following the criterion:

\[
P^m = \{ P_i | | P_i | > m \}
\]

i.e. patterns containing intervals that persists at least for m increasing thresholds. This further selection criterion is related to the height of the shaped bell fragment, in fact a small value of m could represent
are set, defining the expected regions. Finally, the classification is performed by testing the intersection of such regions (see Fig. 3(b)).

First phase of the classification

For each interesting pattern \( P_i \), the dissimilarity \( \delta(P_i, P) \) is evaluated (\( \delta \) is defined in Eq. 4, \( P \) is the model), the rule to classify \( P_i \) is:

\[
c_1(P_i) = \begin{cases} 
L & \text{if } \delta(P_i, P) \leq \phi_1 \\
EW & \text{if } \phi_1 < (P_i, P) \leq \phi_2 \\
ED & \text{otherwise}
\end{cases}
\]  

(5)

where \( L \) means linker pattern, EW or ED are nucleosomal pattern, and in particular expected well positioned patterns and expected delocalized patterns respectively.

Second phase of the classification

Afterwards, for each expected well positioned nucleosomal pattern \( P_i = \{ f_i^1, f_i^2, \ldots, f_i^{l-1}, f_i^l \} \) (e.g. \( c_1(P_i) = EW \)), the center of the nucleosomal region \( C_i \) is calculated:

\[
C_i = \frac{\sum_{j=1}^{l-1} f_i^j + f_i^{l+1}}{2l}
\]

(6)

which represents the mean of the first \( l \) intervals defining the pattern \( P_i \).

Conversely, for each expected delocalized nucleosomal pattern (e.g. \( c_1(P_i) = ED \)), the delocalized interval \( B_i', E_i' \) is defined such that:

\[
B_i' = \frac{1}{l/2} \sum_{j=1}^{j=l/2} b_i^j \quad \text{and} \quad E_i' = \frac{1}{l/2} \sum_{j=1}^{j=l/2} e_i^j
\]

(7)

Note that, \( B_i' \) and \( E_i' \) represent respectively the mean of the first \( l/2 \) beginning and ending of each interval belonging to the pattern \( P_i \). The expected regions is so defined

\[
A_i = \begin{cases} 
\{ C_i(l-3), C_i(l) + 3 \} & \text{if } c_1(P_i) = EW \\
[ B_i', E_i' ] & \text{otherwise}
\end{cases}
\]

(8)

In particular, each expected region \( A_i \) is, in the case \( P_i \) is an expected well positioned pattern, an interval with beginning and ending 3 probes before and ending 3 probes after the center \( C_i \), otherwise it is the interval \( B_i', E_i' \).

Finally, the classification rule is

\[
c_2(P_i) = \begin{cases} 
F & \text{if } A_i \cap A_i' = \emptyset \text{ or } i, \text{ otherwise}
\text{ otherwise}
\end{cases}
\]

(9)

where \( F, W \) and \( D \) stand for fused, well positioned, delocalized nucleosomes respectively (see Fig. 3a). Informally, the classification rule in Equation 9 assign the fused class if the expected nucleosomal regions overlap otherwise confirm the classification of the first phase.

Synthetic signal generation

Before validating the MLM on biological data, a procedure to generate synthetic signal has been developed allowing us to assess the feasibility of our method on controlled data. The model is characterized by several parameters \( \lambda, r, o, n, m, \) \( \alpha \), \( \beta \), \( \gamma \) \( \delta \). Description of which is given in supplementary Table 1S.

Initially, a binary mask signal \( M \) is generated by considering as 1’s all the base pairs representing a nucleosome (the nucleosomal regions) and as 0’s the regions representing linkers (the linker regions). Note that, the beginning of each nucleosomal region is established by the Poisson distribution with mean \( \lambda \). The mask signal \( M \) will be used in order to validate the MLM. The red channel of the microarray (the genomic channel) results from the generation of \( nR \) replicates \( f_1^R, \ldots, f_{nR}^R \) each one starting from an initial nucleosomal region of random size \( b \sim U(0, r) \) (uniformly distributed in the range \( [0, r] \)), followed by continuous nucleosomal region of \( r \) base pairs. Conversely, in order to simulate the green channel (the nucleosomal channel) \( nR \) replicates, \( f_1^R, \ldots, f_{nR}^R \) are considered, each one initially equal to \( M \) and subsequently modified by perturbing each staring points \( x_0^i \) of the nucleosome by random \( \mu \sim U(0, r) \), so that \( x_0^i = x_0^i + \mu \). Note that the percentage of nucleosomes to consider as delocalized is established by the parameter \( dp \). Afterwards, each nucleosomal region on the generic replicate \( f_i^R \) and \( f_i^R \) can be switched off depending on a the value of a random variable \( \alpha \sim U(0, 1) \). Finally, each nucleosomal region verifying the test \( \alpha \text{--pur} \) is considered and set to 1, otherwise it is not considered and set to 0. This results in new replicates \( T^{R}_1 \) and \( T^{R}_2 \).

Finally, the generated synthetic signal is so defined:

\[
V(i) = \left\{ \begin{array}{ll}
\log_2 \left( \sum_{j=1}^{nR} \frac{T_i^{R}(k) \cdot \text{ra}}{T_i^{R}(k)} + \epsilon \right) & \text{if } (r-o) \leq r + 1 \leq (r-o) + o \\
\end{array} \right.
\]

where \( \epsilon \sim N(0,1, \text{nsv}) \).

Parameter selection by calibration

In order to set the proper values of \( K \) (number of thresholds), and \( m \) (the minimum number of permanences), a calibration procedure has been used. In particular, such values has been estimated by studying the plots of particular functions able to measure the goodness of several \( K \) and \( m \).

Estimation of \( m \)

The minimum number of permanences \( m \) has been estimated by using the synthetic signal generator described above. This gives the opportunity to make a massive experimental study on the relation between \( K \) and \( m \). In particular, \( c = 10 \) copies at different signal to noise ratio \( j = 1, 2, 4 \) has been generated, resulting in a total of 3 x 10 synthetic signals \( V_i \). Once fixed a signal to noise ratio \( j \), for each \( V_i \) the value of \( m \) which maximizes the recognition performances for several thresholds for \( k = 20, \ldots, 50 \) has been found.

Supplementary Fig. 1S shows the results performed by considering \( c = 10 \) copies, three signal to noise ratio values \( 1, 2, 4 \), and \( k = 20, \ldots, 50 \) thresholds. In each plot, the x axis represents the number of thresholds \( k \) (i.e. number of cuts), the column bar groups the best recognition (Supplementary Fig. 1S(a)) and the percentage of minimum number of permanences which causes the best performances (Supplementary Fig. 1S(b)) on all the experiments. From this experimental study, it emerges that the use of an high number of thresholds can compromise the recognition process, moreover, the \( m \) value seems not dependent from \( K \), and the one which causes the best recognition ranges in an interval of \( [0.15 \times K, 0.30 \times K] \).

Estimation of \( K \)

The proper value of \( K \) is estimated starting from the convolved input signal \( X \). Giving a convoluted signal fragment \( X_i \), we resample it in the \( y \) direction resulting in several resamples \( X_i^{R} \) for different threshold values \( k = 1, \ldots, K_{\text{max}} \). We can measure the goodness of \( k \) by the average normalized correlation \( \rho(k) \) and the average missing probes \( \text{MS}(k) \) so defined:

\[
\rho(k) = \frac{1}{T_t} \sum_{i=1}^{T_t} \left[ 1 + \rho^2 \left( S_i(S_i^R) \right) \right]
\]

\[
\text{MS}(k) = \frac{1}{T_t} \sum_{i=1}^{T_t} \text{MS}(k, t)
\]

In particular \( \rho(k) \) measures the average normalized correlation between each resample \( X_i^{R} \) and the generic fragment \( X_i \) (\( \rho \) is the pearson correlation coefficient), while \( \text{MS}(k) \) the average of the
missing probe values $MS(k, t)$ due to the resample of $X_k$ by $k$ thresholds. Finally the value $K$ is selected interactively by looking both at the plots of $\varrho$ and $MS$, searching for the best compromise of maximum $\varrho$ and minimum $MS$ (Supplementary Fig. 2S).

Results

The following experiments have been carried out by measuring the correspondence between Nucleosome and linker regions. In the case of the synthetic signal, the output of the classifier has been compared with a mask $M'$ derived from $M$, in the case of the real data set it has been compared with the output of the Hidden Markov model (HMM) used in the paper of Yuan et al. [8] optimally converted into a binary string.

In all the experiments, the same value $(\varphi_1, \varphi_2)=(\text{mean}(\delta(F_t, F_l))\pm 3\text{std}(\delta(F_t, F_l)))$, $\text{mean}(\delta(F_t, F_l))\pm 3\text{std}(\delta(F_t, F_l))$ has been considered, where $F_t$ are all the sub-fragments used on the construction of the model $F$. Moreover, by biological consideration, the radius os has been set to os=4. The performances have been evaluated in terms of Recognition Accuracy, RA. The RA uses a new mask $M'$ obtained by converting $M$ into probe coordinates such that a probe value is set to 1 (e.g. shows a nucleosome portion) if the corresponding base pairs in $M$ include at least a 1. The real nucleosomal (linker) regions RNR (RLR) are represented by $M'$ as contiguous sequence of 1's or 0's respectively, here we consider that a nucleosomal (linker) region C NR (CLR) has been classified correctly if there is a match of at least $l=0.7\times L$ contiguous 1's (0's) between C NR (CLR) and the corresponding RNR (RLR) in $M'$ where $L$ is the length of RNR (RLR). The value 0.7 has been chosen because it represents a 70% of regions overlap very unlikely to be due to chance.

MLM vs HMM on synthetic data

For the calibration phase $K=20$ and $m=5$, the value of in Eq. 4 has been set to 0.5 to equally balance the two component of the dissimilarity. In particular, 6 signal of length ranging from 2337 probes (70130 bp) to 2361 probes (70850 bp) have been generated for the signal to noise ratio values 1, 2, 4, 6, 8 and 10. The other parameters used to generate such signals are reported in supplementary Table 1S. In Fig. 4 the results of the total RA for all the experiments are reported. The confusion matrices of HMM and MLM for all the experiments are reported in supplementary Tables 2S, 3S respectively. In Fig. 4 the results of the total RA for all the experiments is summarized. Fig. 4 shows that the HMM is slightly more accurate in finding the bounds of the nucleosome regions. The synthetic results can be summarized in an overall RA of 0.96 for the MLM and 0.98 for HMM.

Table 1

<table>
<thead>
<tr>
<th>RA</th>
<th>H</th>
<th>L</th>
<th>N</th>
<th>M</th>
<th>H</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>L</td>
<td>0.79</td>
<td>0.21</td>
<td>L</td>
<td>0.52</td>
<td>0.47</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>0.13</td>
<td>0.87</td>
<td>M</td>
<td>N</td>
<td>0.12</td>
</tr>
</tbody>
</table>

The table on the left shows the RA results of HMM when considering MLM as the truth classification, while the opposite is shown on the right table.

MLM vs HMM on real data

In this experiment, we have compared the accordance of the two models on the S. cerevisiae real data. The input signal representing this data is composed by 215 contiguous fragments for a total of 24167 base pairs. In such experiment, we have chosen $K=40$, $m=6$ by the calibration phase ($m=0.15\times 40$) and $\alpha=0.5$ to equally balance the two component of the dissimilarity (see the definition in Eq. 4). The confusion matrices which show the RA of HMM considering MLM as the truth classification and RA of MLM considering HMM as the truth classification are reported in Table 1. The results can be summarized in an overall RA of 0.83 for the HMM (MLM true) and 0.69 for MLM (HMM true). In particular, from this studies we can conclude that MLM does not fully agree with HMM on the linkers patterns. Remarkably, when we compared both MLM and HMM and data coming from recently developed deep sequencing approach (DS) Pugh et al. [22] we found a better agreement with MLM (0.58) rather than with HMM (0.44) (supplementary Table 4S, and supplementary Fig. 3S). These analyses indicate that the integration of the HMM and MLM could improve the overall classification.

Computational notes

The computation times of MLM and HMM have been compared on 10 experiments. In particular, 10 synthetic signals have been generated, each one with a fixed number of well positioned nucleosomes ranging from 10 to 100 by step of 10. In supplementary Fig. 4S, the ratios between the execution time of MLM ($T_m$) and HMM ($T_h$) for each experiment are shown. From this study, it results that, on average, $T_h=1.7\times 10^4\times T_m$.

Discussion and future work

We have developed a new method that can be successfully used to identify genome wide nucleosome positions starting from tiling array

![Fig. 4. Results on synthetic data: The Recognition Accuracy of MLM and HMM on 6 synthetic signals generated at signal to noise ratio 1, 2, 4, 6, 8 and 10.](image-url)
data. We have also defined a method to generate synthetic microarray data fully inspired from the microarray technique that has been used in [8]. However, since MLM can localize nucleosomes based on shape information we expect that our method could be easily extended to the analysis of data coming from newly developed “deep sequencing” approaches. We have tested our method on both synthetic and real data, reaching in the first case a recognition of 96% and in the second case an accuracy of 76% with the Hidden Markov Model with a gain in computation time of ~1.7×10^4 with respect to the latter. The great improvement in computational time of the MLM over standard statistical methods, like HMM, makes the MLM a method of choice for the analysis of genome-wide nucleosome position starting from more complex higher density arrays or very large “deep sequencing” data. Nucleosome spacing and mobility increase in complexity as we move from lower to higher eukaryote genomes. The ability to recognize nucleosomes with different mobility characteristics (well positioned, delocalized, fused) is directly linked to the pattern shape recognition feature integrated into the MLM approach we developed. However, new methods to efficiently map or predict nucleosome positions have been recently developed (see [18,19]). Although, we predict that the MLM method will be particularly suited for the genome wide nucleosome position analysis of complex chromatin present in higher eukaryote model organisms, future work will be necessary to cross compare the efficiency of different nucleosome mapping algorithms.

Acknowledgements

This work makes use of results produced by the PI2S2 Project managed by the Consorzio COMETA, a project co-funded by the Italian Ministry of University and Research (MIUR) within the Piano Operativo Nazionale “Ricerca Scientifica, Sviluppo Tecnologico, Alta Formazione” (PON 2000–2006). More information is available at http://www.pi2s2.it and http://www.consorzio-cometa.it. D.F.V.C. was supported by Fondazione Telethon, Giovanni Armenise Harvard Foundation, MIUR, HFSP and Compagnia San Paolo. G.Y.’s research was funded by Dana-Farber Cancer Institute and the Claudia Adams Barr Program.

Appendix A. Supplementary data


References