Genomics 93 (2009) 140–145

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

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 Dulbecco, 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.

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, the 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

* Corresponding author. Fax: +39 0916040311.
E-mail address: lobosco@math.unipa.it (G. Lo Bosco).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.09.012

task of peak detection remains a key problem for the statistical analysis of the input data. Unlike microarray-based approaches, where data collection is constraint 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 folding [26] and the positioning of nucleosomes along chromatin 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 underlie 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_i (*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 subsamples 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.

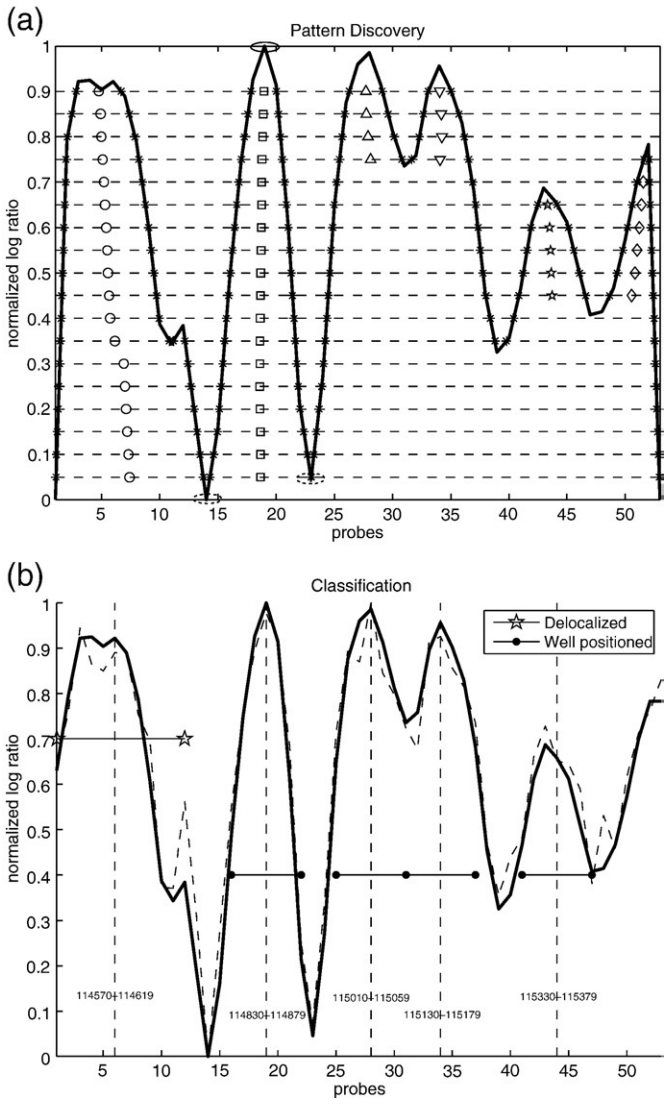

Fig. 1. (a) Input signal, smoothing, pattern identification and extraction: A *Saccharomyces cerevisiae* microarray data portion. Each x value represents a spot (probe) on the microarray and the corresponding y value is the logarithmic ratio of its Green and Red values. Nucleosomes regions are around the peaks signal (one is marked by black circle), while lower ratio values show linker regions (marked by dashed circles). The dashed lines represents the threshold levels, in this example 6 patterns are retrieved, identified by rhombus, circle, square, triangle down, triangle up, star. Each pattern identifier is replicated for each of its feature values and pointed in each one of its middle point. (b) An example of classification: In this portion 5 nucleosome regions are shown together with its range in base pairs. In particular 1 out of the 5 regions is classified as *delocalized* while the remaining *well positioned*.

MLM preprocessing step

A preprocessing is necessary in order to reduce the effect of the signal noise. Each fragment S_t, $1 \le t \le T$ 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* $w = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right]$ [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

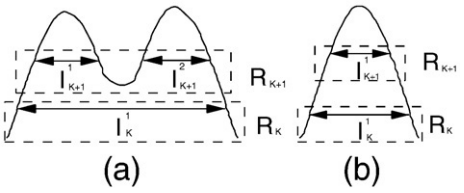

Fig. 2. Two different shapes of the input signal: (a) Since at threshold level $K+1$ the interval $R_k=\{I_K^1\}$ has two subset $R_{k+1}=\{I_{K+1}^1, I_{K+1}^2\}$, we set three pattern $P_1=\{I_K^1\}$, $P_2=\{I_{K+1}^1\}$ and $P_3=\{I_{K+1}^2\}$. (b) In this case, I_{K+1}^1 is the unique subset of I_K^1, thus we set an unique pattern $P_1=\{I_K^1, I_{K+1}^1\}$.

are opportunely selected for the model definition. Each convolved fragment X_t is processed in order to find $L(X_t)$ local maxima $M_t^{(l)}$ for $l=1,\ldots, L(X_t)$. The extraction of each sub-fragment for each $M_t^{(l)}$ is performed by assigning all values in a window of radius centered in $M_t^{(l)}$ to a vector, *F_t^l of size $2\times os+1$: $F_t^l(j)=X_t(M_t^{(l)}-os+j-1)$, for $j=1, 2,\ldots, 2\times os+1$.* The selection process extracts the significant sub-fragments to be used in the model definition. This is performed by satisfying the following rule:

$$\begin{cases} F_t^l(j+1)-F_t^l(j) > 0 & j=1,\ldots,os \\ F_t^l(j+1)-F_t^l(j) < 0 & j=os+1,\ldots,2\times os \end{cases} \quad (1)$$

After this selection process $G(X_t)$ sub-fragments remain for each X_t. The model of the *interesting pattern* is then defined by considering the following average:

$$\overline{F}(j)=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{G(X_t)}\sum_{k=1}^{G(X_t)}F_t^k(j)\quad j=1,\ldots,2\times os+1 \quad (2)$$

That is, for each j, the average value of all the sub-fragments satisfying Eq. 1.

MLM interval identification step1

The core of the method is the interval identification by considering K threshold levels t_k ($k=1,\ldots, K$) of the convolved signal **X**. For each t_k a set of intervals $R_k=\{I_k^1, I_k^2, \ldots, I_k^{n_k}\}$ is obtained; where, *$I_k^i=[b_k^i, e_k^i]$ and $X1(b_k^i)=X(e_k^i)=t_k$.* In Parameter selection by calibration section a calibration procedure to select the proper value of K is described.

MLM interval merging and pattern definition step

This step is performed by taking in account that bell shaped pattern must be extracted for the classification phase. Such kind of patterns are characterized by sequences of intervals $\{I_j^1, I_{j+1}^2,\ldots, I_{j+1}^n\}$ such that $I_j^i \supseteq I_{j+1}^{i+1}$; more formally a pattern P_i is defined as:

$$P_i=\left\{I_j^{i_j}, I_{j+1}^{i_{j+1}},\ldots, I_{j+l}^{i_{j+l}} \mid \forall I_k^{i_k}\exists! I\in R_{k+1} : I=I_{k+1}^{i_{k+1}}\subseteq I_k^{i_k}\right\}$$

where, j defines the threshold, t_j, of the widest interval of the pattern. From the previous definition it follows that P_i is build by adding an interval $I_{k+1}^{i_{k+1}}$ only if it is the unique in R_{k+1} that is included in $I_k^{i_k}$. 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* $\mathbf{P^{(m)}}$ are selected following the criterium:

$$\mathbf{P^{(m)}}=\{P_i : |P_i| > m\} \quad (3)$$

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 represents 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 $P_i\in\mathbf{P^{(m)}}$ is identified by $I_j^{i_j}, I_{j+1}^{i_{j+1}},\ldots, I_{j+l}^{i_{j+l}}$, 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:

$$\delta(P_r,P_s)=(1-\alpha)(A_r-A_s)+\alpha\sum_{i\in I}\left(a_i^{r_i}-a_i^{s_i}\right) \quad (4)$$

where, A_r and A_s are the surfaces of the two polygons bounded by the set of vertexes $V=\cup_{i\in I}\left\{\left(b_i^{r_i}, e_i^{r_i}\right), \left(b_i^{s_i}, e_i^{s_i}\right)\right\}$, $a_i^{r_i}=e_i^{r_i}-b_i^{s_i}$, $a_i^{s_i}=e_i^{s_i}-b_i^{s_i}$, and α 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 include 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

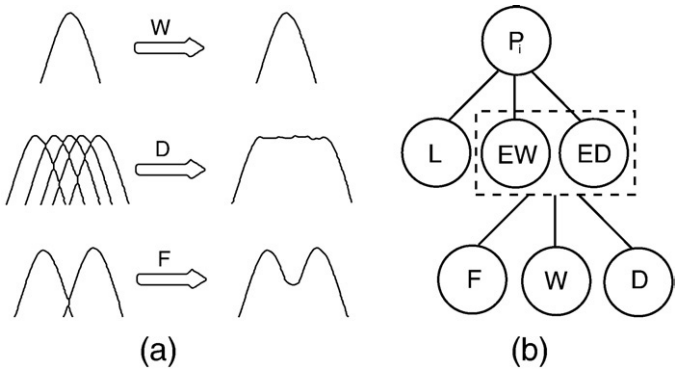

Fig. 3. (a) Shapes of the patterns: The three classes of nucleosomes we can detect with the MLM very likely reflect different nucleosome mobility existing in vivo at specific chromatin loci. Delocalized nucleosomes probably represent single nucleosomes or arrays of nucleosomes with high mobility, while fused nucleosomes may reflect a single nucleosome that occupies two distinct close positions in different cells. On the left of the arrows, the particular nucleosome configuration which generates the resulting shape of well positioned (*W*), delocalized (*D*) and fused (*F*) nucleosome classes are shown. (b) Classification: The classification of a generic pattern P_i is performed into two phases. In the first phase the linker (*L*), the expected well positioned (*EW*) and the expected delocalized (*ED*) patterns are established by using the classification rule defined by c_1. In the second phase, the expected regions A_i are defined by opportunely processing *EW* and *ED* patterns, and afterwards used by the classification rule c_2 in order to finally classify between well positioned (*W*), delocalized (*D*) and fused (*F*) nucleosomes.

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 $\delta(P_i,\overline{F})$ is evaluated (δ is defined in Eq. 4, $\overline{F}$ is the model), the rule to classify P_i is:

$$c_1(P_i)=\begin{cases} L & \text{if } \delta(P_i,\overline{F})\le\phi_1 \\ EW & \text{if } \phi_1<(P_i,\overline{F})\le\phi_2 \\ ED & \text{otherwise}\end{cases} \tag{5}$$

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 $Pi=\left\{I_j^i, I_{j+1}^{i_{j+1}}, \ldots, I_{j+l}^{i_{j+l}}\right\}$ (e.g. $c_1(P_i)=EW$), the *center of the nucleosomal region* C_i is calculated:

$$Ci=\frac{1}{l}\sum_{k=j}^{j+l}\frac{e_k^i+b_k^i}{2} \tag{6}$$

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:

$$B^i=\frac{1}{l/2}\sum_{k=j}^{j+(l/2)}b_k^i \text{ and } E^i=\frac{1}{l/2}\sum_{k=j}^{j+(l/2)}e_k^i \tag{7}$$

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

$$A_i=\begin{cases}[C_i(l)-3, C_i(l)+3] & \text{if } c_1(P_i)=EW \\ [B^i,E^i] & \text{otherwise}\end{cases} \tag{8}$$

In particular, each expected region A_i is, in the case P_i is an expected well positioned pattern, an interval with beginning 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

$$c_2(P_i)=\begin{cases} F \text{ if } A_i\cap A_j\neq\varnothing j\neq i, \text{otherwise} \\ \left[\begin{array}{l} W \text{ if } c_1(P_i)=EW \\ D \text{ if } c_1(P_i)=ED\end{array}\right.\end{cases} \tag{9}$$

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 (*nn*, *nl*, λ, *r*, *o*, *nr*, *dp*, *dr*, *nsv*, *pur*, *ra* and *snr*), 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 λ. 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 $I_1^R,\ldots,I_{nr}^R$ 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 nucleosomic region of *r* base pairs. Conversely, in order to simulate the green channel (the nucleosomic channel) *nr* replicates, $I_1^G,\ldots,I_{nr}^G$ are considered, each one initially equal to *M* and subsequently modified by perturbing each staring points x_D^i of the nucleosome by random $\mu\sim U(dr)$, so that $x_D^i=x_D^i+\mu$. Note that the percentage of nucleosomes to consider as delocalized is established by the parameter *dp*. Afterwards, each nucleosomic region on the generic replicate I_i^R *and* I_i^G can be switched off depending on a the value of a random variable $\alpha\sim U(0, 1)$. Precisely, each nucleosomal region verifying the test $\alpha<pur$ is considered and set to 1, otherwise it is not considered and set to 0. This results in new replicates T_i^R *and* T_i^G. Finally, the generated synthetic signal is so defined:

$$V(i)=\left\{\log_2\left(\sum_{j=1}^{nr}\frac{T_j^G(k)*ra}{T_j^R(k)}+\varepsilon\right)\middle|(r-o)i-r+o+1\le k\le(r-o)i+o\right\} \tag{10}$$

where $\varepsilon\sim N(0.1, nsv)$.

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 3×10 synthetic signals V_{ij}. Once fixed a signal to noise ratio *j*, for each V_{ij} 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_t we resample it in the *y* direction resulting in several resamples $X_t^{(k)}$ for different threshold values $k=1,\ldots, K_{max}$. We can measure the goodness of *k* by the *average normalized correlation* $\overline{\varrho(\boldsymbol{k})}$ and the *average missing probes* $\overline{MS(k)}$ so defined:

$$\overline{\varrho(\boldsymbol{k})}=\frac{1}{T}\sum_{t=1}^{T}\frac{1+\rho^2\left(S_t,S_t^{(k)}\right)}{2} \tag{11}$$

$$\overline{MS(k)}=\frac{1}{T}\sum_{t=1}^{T}MS(k,t)$$

In particular $\overline{\varrho(\boldsymbol{k})}$ measures the average normalized correlation between each resample $X_t^{(K)}$ and the generic fragment X_t (ρ is the pearson correlation coefficient), while $\overline{MS(k)}$ the average of the

missing probe values $MS(k, t)$ due to the resample of X_t by k thresholds. Finally the value K is selected interactively by looking both at the plots of ϱ and $\overline{MS}$, searching for the best compromise of maximum ϱ and minimum $\overline{MS}$ (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 $(\phi_1, \phi_2)=(mean(\delta(F_t^l, \overline{F}))-3std(\delta(F_t^l, \overline{F})), mean\ (\delta(F_t^l, \overline{F}))+3std(\delta(F_t^l, \overline{F}))$ has been considered, where F_t^l are all the sub-fragments used on the construction of the model $\overline{F}$. 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 nulceosome 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 nulceosomal (linker) region *CNR* (*CLR*) has been classified correctly if there is a match of at least $l=0.7\times L$ contiguous 1's (0's) between *CNR* (*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 *MLM*, we have chosen by 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
Agreement between the *HMM* and *MLM* (and vice versa) on the *Saccharomyces cerevisiae* data set for nucleosomes (N) and linker (L) regions

	M	*L*	*M*		*H*	*M*	*M*
H		*L*	*N*	*M*		*L*	*N*
M	*L*	0.79	0.21	*L*	*L*	0.52	0.47
M	*N*	0.13	0.87	*M*	*N*	0.12	0.87

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 disssimilarity (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

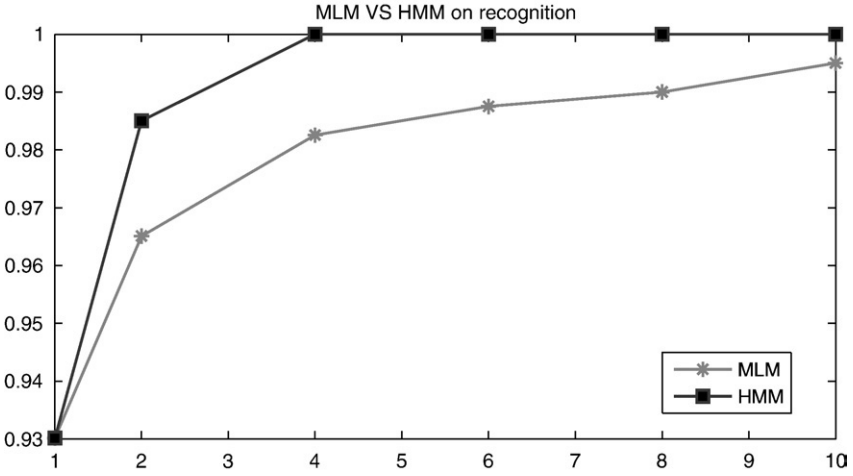

Fig. 4. Results on synthetic data: The Recognition Accuracy of *MLM* and *HMM* on 6 synthetic signals generated at signal to noise ratio 1, 2, 4, 6, 8 and 10.

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 accordance of 76% with the Hidden Markov Model with a gain in computation time of $\sim 1.7 \times 10^4$ 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

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.09.012.

References

[1] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond, Crystal structure of the nucleosome core particle at 2.8 A resolution, Nature 389 (6648) (1997) 251–260.
[2] D.F.V. Corona, J.W. Tamkun, Multiple roles for ISWI in transcription, chromosome organization and DNA replication, Biochim. Biophys. Acta 1677 (1–3) (2004) 113–119.
[3] J. Svaren, W. Horz, Transcription factors vs. nucleosomes: regulation of the PHO5 promoter in yeast, Trends Biochem. Sci. 22 (1997) 93–97.
[4] W. Stunkel, I. Kober, K.H. Seifart, A nucleosome positioned in the distal promoter region activates transcription of the human U6 gene, Mol. Cell. Biol. 17 (1997) 4397–4405.
[5] B.E. Bernstein, C.L. Liu, E.L. Humphrey, E.O. Perlstein, S.L. Schreiber, Global nucleosome occupancy in yeast. Genome Biol. 5 (2004) R62.
[6] D.K. Pokholok, C.T. Harbison, S. Levine, et al., Genome-wide map of nucleosome acetylation and methylation in yeast, Cell 122 (4) (2005) 517–527.
[7] C.K. Lee, Y. Shibata, B. Rao, B.D. Strahl, J.D. Lieb, Evidence for nucleosome depletion at active regulatory regions genome-wide, Nat. Genet. 36 (2004) 900–905.
[8] G.-C. Yuan, Y.J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, O.J. Rando, Genome-scale identification of nucleosome positions in *S. cerevisiae*, Science 309 (2005) 626–630.
[9] C.T. Harbison, D.B. Gordon, T.I. Lee, et al., Transcriptional regulatory code of a eukaryotic genome. Nature 431 (2004) 99–104.
[10] W. Lee, D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, C. Nislow, A high-resolution atlas of nucleosome occupancy in yeast, Nat. Genet. 39 (10) (2007) 1235–1244.
[11] I. Whitehouse, O.J. Rando, J. Delrow, T. Tsukiyama, Chromatin remodelling at promoters suppresses antisense transcription, Nature 450 (7172) (2007) 1031–1035.
[12] M. Buck, B.A. Nobel, J.D. Lieb, ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data, Genome Biol. 6 (11) (2005) R97.
[13] H. Ji, W.H. Wong, TileMap: create chromosomal map of tiling array hybridizations, Bioinformatics 21 (18) (2005) 3629–3636.
[14] T.H. Kim, B. Ren, Genome-wide analysis of protein–DNA interactions, Annu. Rev. Genomics Hum. Genet. 7 (2006) 81–102.
[15] Z.D. Zhang, J. Rozowsky, H.Y.K. Lam, J. Du, M. Snyder, M. Gerstein, Tilescope: online analysis pipeline for high-density tiling microarray data, Genome Biol. 8 (5) (2007) R81.
[16] W.E. Johnson, W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown, X.S. Liu, Model-based analysis of tiling-arrays for ChIP-chip, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 12457–12462.
[17] S. Keles, M.J. Van Der Laan, S. Dudoit, S.E. Cawley, Multiple testing methods for ChIP-Chip high density oligonucleotide array data, J. Comput. Biol. 13 (3) (2006) 579–613.
[18] V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes, T. Grange, DNA physical properties determine nucleosome occupancy from yeast to fly, Nucleic Acids Res. 36 (11) (2008) 3746–3756.
[19] M. Yassour, T. Kaplan, A. Jaimovich, N. Friedman, Nucleosome positioning from tiling microarray data, Bioinformatics 24 (13) (2008) i139–i146.
[20] F. Ozsolak, J.S. Song, X.S. Liu, D.E. Fisher, High-throughput mapping of the chromatin structure of human promoters, Nat. Biotechnol. 25 (2007) 244–248.
[21] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J.P. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[22] I. Albert, T.N. Mavrich, L.P. Tomsho, J. Qi, S.J. Zanton, S.C. Schuster, B.F. Pugh, Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome, Nature 446 (2007) 572–576.
[23] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome, Cell 129 (2007) 823–837.
[24] T.S. Mikkelsen, M. Ku, D.B. Jaffe, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (2007) 553–560.
[25] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions, Science 316 (2007) 1497–1502.
[26] A.L. Delcher, S. Kasif, H.R. Goldberg, W.H. Hsu, Protein secondary structure modelling with probabilistic networks, Proc. of Int. Conf. on Intelligent Systems and Molecular Biology, 1993, pp. 109–117.
[27] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer, 2001.
[28] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inf. Theory 48 (2002) 1518–1569.
[29] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Stat. 37 (1983) 36–48.
[30] R.G. Lyons, Understanding Digital Signal Processing, Addison Wesley, 1997.