BIOINFORMATICS

Vol. 00 no. 00
Pages 1–7

In vitro vs *in vivo* compositional landscapes of histone sequence preferences in eucaryotic genomes

Raffaele Giancarlo [1,†], Simona E. Rombo [1,†*] and Filippo Utro [2,†]

[1]Dipartimento di Matematica ed Informatica, Università degli Studi di Palermo, Via Archirafi 34, 90123 Palermo, Italy

[2]Computational Biology Center, IBM T. J. Watson Research, Yorktown Heights, NY 10598, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Although the nucleosome occupancy along a genome can be in part predicted by *in vitro* experiments, it has been recently observed that the chromatin organization presents important differences *in vitro* with respect to *in vivo*. Such differences mainly regard the hierarchical and regular structures of the nucleosome fiber, whose existence has long been assumed, and in part also observed *in vitro*, but that does not apparently occur *in vivo*.

It is also well known that the DNA sequence has a role in determining the nucleosome occupancy.

Therefore, an important issue is to understand if, and to what extent, the structural differences in the chromatin organization between *in vitro* and *in vivo* have a counterpart in terms of the underlying genomic sequences.

Results: We present the first quantitative comparison between the *in vitro* and *in vivo* nucleosome maps of two model organisms (*S. cerevisiae* and *C. elegans*). The comparison is based on the construction of weighted k-mer dictionaries. Our findings show that there is a good level of sequence conservation between *in vitro* and *in vivo* in both the two organisms, in contrast to the abovementioned important differences in chromatin structural organization. Moreover, our results provide evidence that the two organisms predispose themselves differently, in terms of sequence composition and both *in vitro* and *in vivo*, for the nucleosome occupancy. This leads to the conclusion that, although the notion of a genome *encoding* for its own nucleosome occupancy is general, the intrinsic histone k-mer sequence preferences tend to be species-specific.

Availability: The files containing the dictionaries and the main results of the analysis are available at `http://math.unipa.it/rombo/material`.

Contact: {raffaele.giancarlo,simona.rombo}@unipa.it, futro@us.ibm.com.

*To whom correspondence should be addressed.

†All three of the authors have to be regarded as joint First Author.

1 INTRODUCTION

The nucleosome fiber, in which DNA is wrapped around core histones, has long been assumed to be folded according to several hierarchical levels: starting from a 10-nm chromatin fiber, it would be then packed into a 30-nm fiber, and further helically folded in a larger fiber to form highly condensed chromosomes (Alberts *et al.*, 2002). However, recent studies show that there are important differences in the organization of chromatin as observed *in vitro* with respect to *in vivo*, especially with reference to these hypothesized hierarchical levels. In particular, although the 30-nm chromatin fiber can be reconstructed *in vitro* (Robinson *et al.*, 2006), it has been elusive to be observed *in vivo* (Tremethick, 2007). As a matter of fact, several findings (Ricci *et al.*, 2015; Hansen, 2012; Razin and Gavrilov, 2014) strongly argue against the existence of a well-organized and ordered fiber *in vivo*, leading to the most recent view that chromosome-level condensation is achieved through packaging of the 10-nm fibers in a fractal manner. Additionally, the accessibility of DNA in chromatin seems to depend on the local mobility of nucleosomes, rather than on decompaction of chromosome regions, which was instead hypothesized in the past (Razin and Gavrilov, 2014).

In this scenario, it is worth to point out that the *in vitro* nucleosome occupancy along a genome, i.e., the *in vitro* reconstruction of the 10-nm fiber, can even be a good predictor of what happens *in vivo*. That is, the *intrinsic* histone DNA sequence preferences in eucaryotic genomes play a role in the determination of nucleosome occupancy *in vivo* (Kaplan *et al.*, 2009). However, although that result is a cornerstone of chromatin studies, a detailed account of the specific changes in histone sequence preferences between *in vitro* and *in vivo* is not yet available.

To clarify the above point, it is worth to recall that, in living cells, nucleosome organization is the result of the concurrent effect of the action of multiple "players", such as chromatin remodellers and site-specific DNA-binding proteins, all competing with histones to bind to their "preferred" DNA sequences (Li *et al.*, 2007; Struhl and Segal, 2013; Tompitak *et al.*, 2017). Therefore, *in vivo*, nucleosome occupancy

maps show the end result of that competition, making it difficult to establish to which extent each specific "player" alters the intrinsic nucleosome organization "encoded" by the underlying genomic sequence. That is, based on an *in vivo* nucleosome map, it is nearly impossible to infer a qualitative/quantitative account of the changes that have involved the genomic positions histones would have chosen, in absence of competition, and based on their DNA sequence preferences only. Therefore, a comparison between *in vitro* and *in vivo* maps is required, possibly involving the known DNA sequence binding affinities of transcription factors. In this respect, a first level of detail is provided by Charoensawan *et al.* (2012) for *S. cerevisiae* and by Locke *et al.* (2013) for *C. elegans*. The first study focuses on the sequence competition between histones and transcription factors in *S. cerevisiae* only, while the second concentrates on a high level comparison between *in vivo* and *in vitro* nucleosome maps in *C. elegans* only, involving to some extent also sequence preferences.

Here we contribute to advance the State of the Art regarding *in vitro* vs *in vivo* histone sequence preferences in several directions, by focusing on two model organisms, i.e., *S. cerevisiae* and *C. elegans*. This research naturally continues our previous studies on the role of the sequence in nucleosome positioning, which were based on *in vivo* nucleosome maps only (Giancarlo *et al.*, 2015; Utro *et al.*, 2016; Giancarlo *et al.*, 2018). In particular, we present a framework for the generation of weighted k-mer dictionaries which allow for a unified view of data coming from different sources, and apply it in order to contribute along two main directions. The first is to study how much of the intrinsic, i.e., *in vitro*, histone sequence preferences are detained *in vivo* within the same species. This provides a much needed additional level of detail with respect to (Locke *et al.*, 2013), as far as worm is concerned, and a novel level of detail regarding yeast. The second contribution is a comparative analysis between yeast and worm, with reference to their nucleosome sequence compositional landscapes *in vitro* and *in vivo*. To the best of our knowledge, such a comparison has not yet been considered, even at a high level, in the Literature.

Our study shows that:

- Despite the important differences in chromatin organization, there is a good histone k-mer preference conservation between *in vitro* and *in vivo* in both the considered organisms.
- The two considered organisms predispose themselves differently, in terms of their histone k-mer preferences, to their *intrinsic*, i.e., *in vitro*, nucleosome occupancy.

In more detail, our findings bring to light that, in both organisms, chromatin has a sequence compositional organization including at least two families of k-mers which significantly characterize nucleosome enrichment/depletion.

The first family is made of k-mers which have high frequency of occurrence and are strongly conserved between *in vitro* and *in vivo*. Those k-mers mainly include poly(dA:dT) and, surprisingly, they have different roles *in vitro* in the two species. Indeed, they are associated to genomic regions which disfavour nucleosome formation in *S. cerevisiae* and favour it in *C. elegans*, contrary to the indication that their stiffness makes them the hallmark of nucleosome depletion (Segal and Widom, 2009). The second family of k-mers are characterized by a very low frequency of occurrence and are responsible of the main differences between *in vitro* and *in vivo* in both organisms. In yeast, this family seems to be involved in favouring the accessibility of DNA in chromatin.

In conclusion, the macroscopic differences observed in laboratory on the three-dimensional chromatin folding have as a counterpart only slight changes in the corresponding DNA sequence. Such changes involve the more compositionally hetherogeneous regions, and this is related to previous results where it has been shown that sequence complexity may influence nucleosome positioning (Utro *et al.*, 2016).

2 MATERIALS AND METHODS

In our study, we have used datasets well established in the Literature. Details on their description are provided in Section 1 of the Supplementary Material, as well as an outline of the associated procedures of relevance for this research. Here we mention only those facts which are important for the full understanding of the remaining part of this manuscript. In particular, we have used both *in vitro* and *in vivo* nucleosomal maps for yeast and worm. According to Kaplan *et al.* (2009), a *nucleosome enriched* (*depleted*, resp.) region is a maximal consecutive region, longer than 50bp, such that each base-pair is (not, resp.) covered by a nucleosome, i.e., its normalized occupancy value is above (below, resp.) its genomic average. The "affinity" between DNA-binding proteins and 8-mers is quantified by the PBM enrichment score (E-score) (Berger and Bulyk, 2006). The classifications in (Fuxman Bass *et al.*, 2016; Charoensawan *et al.*, 2012; Consortium, 2017) have been used in order to assign a biological function to the considered DNA-binding proteins (i.e., transcripion factors and chromatin remodellers).

In the remaining part of this section, the methodology adopted for our analysis is presented. In the following, we refer to the DNA-binding proteins considered here (i.e., transcripion factors and chromatin remodellers) simply as *proteins*.

2.1 A unified framework for the generation of k-mer dictionaries

K-mer dictionaries are a standard tool for the compositional analysis of biological sequences (Giancarlo *et al.*, 2014), with applications in genomics, proteomics (Grabherr *et al.*, 2011; Zhbannikov *et al.*, 2013) and also epigenomics (Giancarlo *et al.*, 2015). Here we generalize the notion of k-mer dictionaries in order to integrate information coming from different data sources (e.g., genomics, epigenomics, *in vitro* or *in vivo* experiments, etc.). To this aim, we propose a unified framework which defines different families of dictionaries,

together with a set of operations they support. Those latter can be used to combine the information stored in the dictionaries, for the extraction of novel knowledge from them.

In the following, we first describe the general framework, then some dictionary specializations, and finally the methodology for dictionary construction.

2.1.1 General framework It is based on the following definition.

DEFINITION 1. *Weighted k-mer dictionary.* A *weighted k-mer dictionary* is a set D of tuples $\langle w, a(w), s(w)\rangle$ such that:

- w is a word of lenght k, i.e., a k-mer;
- $a(w)$ is a list of attributes of w;
- $s(w)$ is a list of weights associated to w.

Although several operations could be defined on the weighted k-mer dictionaries, only the following two are useful within the research presented here.

DEFINITION 2. *Selection.* Let $D = \langle w, a(w), s(w)\rangle$ be a weighted k-mer dictionary and let $c(w)$ be a list of specific conditions (e.g., regular expressions, attribute values, etc.) to be satisfied by records in D. The result of the *selection* applied on D w.r.t. $c(w)$, denoted by $\sigma_{c(w)}(D)$, is a new dictionary D' obtained by deleting from D all of its records where $c(w)$ is not satisfied.

DEFINITION 3. *Intersection.* Let $D_1 = \langle w, a_1(w), s_1(w)\rangle$ and $D_2 = \langle w, a_2(w), s_2(w)\rangle$ be two weighted k-mer dictionaries. Let w_x and w_y be two k-mers stored in D_1 and D_2, respectively, and such that $w_x = w_y$ (i.e., they are the same k-mer). Then, let $a'(w_x) = a_1(w_x) \cup a_2(w_y)$ and $s'(w_x) = s_1(w_x) \cup s_2(w_y)$. *The result of the* intersection *between D_1 and D_2, denoted by $D_1 \cap D_2$, is the new dictionary* $D' = \langle w, a'(w), s'(w)\rangle$.

2.1.2 Dictionary specializations A first family of dictionaries we present is useful to single out k-mers such that their frequency of occurrence is *statistically significant* in order to characterize one between two input sequence datasets (which are supposed to characterize two different experimental and/or biological conditions). We need two preliminary definitions.

DEFINITION 4. *Empirical probability distribution.* Let w be a k-mer of lenght $\widehat{k}$. Let $f(w)$ be the frequency of occurrence of w in a set of sequences S. Let n be the sum of the frequencies of all k-mers of lenght $\widehat{k}$ occurring in S. The empirical probability of w w.r.t. S is $p(w) = f(w)/n$.

DEFINITION 5. *Z-score.* Let w be a k-mer of lenght $\widehat{k}$. Let S_1 and S_2 be two sets of sequences of arbitrary length on an alphabet Σ. Let $p_1(w)$ and $p_2(w)$ be the empirical probabilities of w w.r.t. S_1 and S_2, respectively. The z-score of w is defined as: $z(w) = \frac{d(w)-avg}{\sqrt{var}}$, where $d(w) = |p_1(w) - p_2(w)|$, $avg = \sum_w d(w)/|\Sigma|^{\widehat{k}}$ and $var = \sum_w (d(w) - avg)^2$.

It is worth to point out that, according to our study, the value of the z-score does not measure the statistical significance of a k-mer *per se*. Instead, it gives an idea of how much the difference between its frequency of occurrence in the two sets S_1 and S_2 deviates from the mean, compared to the other k-mers of the same length.

DEFINITION 6. *Significance dictionary.* Let S_1 and S_2 be two sets of sequences and let h be a real value. A *significance dictionary* is a weighted k-mer dictionary D_S such that $a(w) = \langle k, S\rangle$ and $s(w) = z(w)$, where k is the lenght of w, S identifies the dataset S_x such that $p_x(w) > p_y(w)$ $(x, y \in \{1, 2\})$, and $z(w) > h$.

According to the different ways to choose the threshold value h, different types of significance dictionaries may be constructed. We discuss the specific choice adopted here in Section 2.1.3.

The following family of dictionaries stores k-mers which have high affinity with specific biochemical structures.

DEFINITION 7. *Affinity dictionary.* An *affinity dictionary* is a weighted k-mer dictionary D_A such that, for a given k-mer w, $a(w) = \langle p, F(p)\rangle$ and $s(w) = e(w)$, where p identifies a cellular component (e.g., a protein), $F(p)$ identifies its function and $e(w)$ is an affinity score between w and p, e.g., binding strength.

2.1.3 Dictionary construction and analysis For both yeast and worm, the significance dictionary associated to each of the available maps has been generated, *in vitro* and *in vivo*, respectively, as outlined next. In particular, each map is used to obtain two sets S_1 and S_2, which contain sequences associated to nucleosome enriched or depleted regions, respectively. Those regions are extracted from each map according to the procedure outlined in Section 1.2 of the Supplementary Material. Σ coincides with the alphabet of the four nucleic acids. Suitable values of k in this context are $k = 1 \ldots 9$, as discussed by Giancarlo *et al.* (2015). As for the choice of h in Definition 6, we consider the following procedure.

Let $S = S_1 \cup S_2$. Shuffle S for a number $\widehat{n}$ of times and let $S'(i)$ be the resulting set of sequences at each iteration i ($i = 1, \ldots, \widehat{n}$). Let $S'_1(i)$ and $S'_2(i)$ be two sets of sequences such that $S'(i) = S'_1(i) \cup S'_2(i)$ and $|S'_1(i)| = |S_1|$ and $|S'_2(i)| = |S_2|$, respectively. For each k-mer w, let $z_i(w)$ be the z-score of w w.r.t. $S'_1(i)$ and $S'_2(i)$ at iteration i. Then h is set equal to the maximum value of $z_i(w)$. This correponds to a significance level less than 1% with Bonferroni Correction.

Intersection has been applied to those dictionaries coming from different maps, which are available for the same organism in the same case (e.g., worm *in vivo*). Four different statistical dictionaries, that we refer to as *epigenomic dictionaries* in the following, have been obtained this way: DY_{VT} for yeast *in vitro*, DY_{VV} for yeast *in vivo*, DW_{VT} for worm *in vitro*, and DW_{VV} for worm *in vivo*.

Starting from the epigenomic dictionaries just introduced above, additional dictionaries have been built in order to

investigate the role of homopolymeric tracts in nucleosome formation, and their possibly different role between *in vitro* and *in vivo*. To this aim, the following definition is needed in order to set a specific condition for the selection of records containing homopolymeric tracts from the epigenomic dictionaries.

DEFINITION 8. *Poly(dX:dY) tract.* Let w be a k-mer and let n_{XY} be the number of letters X or Y that w contains, respectively. Then w is a poly(dX:dY) tract if one of the following cases holds, which empirically account for the presence of a core of consecutive identical letters:

- w contains only X (or only Y), if $k = 2, \ldots, 4$.
- w contains four consecutive X (or Y), if $k = 5$.
- w contains four consecutive X (or Y), and $n_{XY} \geq 5$, if $k = 6, 7$.
- w contains four consecutive X (or Y), and $n_{XY} \geq 6$, if $k = 8, 9$.

For each epigenomic dictionary, selection has been applied by setting the condition $c(w) = w$ *is a poly(dA:dT) or a poly(dC:dG) tract*, respectively. This lead to the construction of four *homopolymeric tracts dictionaries*: HDY_{VT}, HDY_{VV}, HDW_{VT} and HDW_{VV}.

As for the affinity dictionaries, one has been constructed for yeast and one for worm, storing in both cases the 8-mers that show high DNA sequence binding affinity with some proteins, measured by PBM experiments, as well as the 9-mers containing at least one of these 8-mers. In such affinity dictionaries, p is a protein associated to w, $F(p)$ is one among Activators (A), Remodelers (C), Dual (D), Repressors (R), Unknown (U), according to (Fuxman Bass *et al.*, 2016; Charoensawan *et al.*, 2012; Consortium, 2017), and $e(w)$ is the E-score between w and p, which quantifies the relative binding preference of a protein as explained in Section 1 of the Supplementary Material (in the case of 9-mers, the minimum E-score of the contained 8-mers is reported). Selection has been then applied under the condition $e(w) \geq 0.45$. Intersection between the resulting dictionaries and the epigenomic dictionaries (in corresponding cases, e.g., yeast *in vitro*) has been computed. The four obtained dictionaries, called *context dictionaries* and denoted by CDY_{VT}, CDY_{VV}, CDW_{VT} and CDW_{VV}, have been further processed in order to quantify whether or not a protein is mainly associated to nucleosome enriched or depleted regions. To this aim, let $\widehat{D}$ be one of such dictionaries and $\widehat{p}$ be a protein in $\widehat{D}$. For each record $\widehat{r}$ in $\widehat{D}$ that contains $\widehat{p}$, let n_1 (n_2, resp.) be the number of k-mers in $\widehat{r}$ which characterize nucleosome depleted (enriched, resp.) regions in terms of a sequence context, as explained next. If $n_1 > n_2$, then $\widehat{p}$ is associated to a larger number of k-mers which characterize depletion, i.e., it has a *context of depletion*; if $n_1 < n_2$, it has a *context of enrichment*. In case of ties, $\widehat{p}$ has a *neutral context*.

3 RESULTS

Section 3.1 is devoted to describe the analysis of the epigenomic dictionaries. Section 3.2 describes the analysis of the homopolymeric tracts dictionaries. In Section 3.3 the analysis of the context dictionaries is presented.

3.1 Analysis of the epigenomic dictionaries

3.1.1 Conservation of histone sequence preferences: in vitro vs in vivo The first study described here aims to show to what extent the intrinsic, i.e., *in vitro*, histone k-mer preferences are preserved *in vivo*. To this end, Table 1 reports the number of k-mers, and the percentage of k-mers characterizing nucleosome enriched/depleted regions (denoted by $+/-$, respectively), stored in the epigenomic dictionaries. The same statistics are reported also for the intersection between *in vitro* and *in vivo* dictionaries, for both *S. cerevisiae* and *C. elegans*.

From a first level analysis, it is evident that there is a good k-mer histone preference conservation (61% and 72% in yeast and worm, respectively). With reference to the important differences in chromatin organization observed between *in vitro* and *in vivo* (Ricci *et al.*, 2015), such a degree of conservation is not obvious.

A comparative analysis between yeast and worm (based on Table 1) shows that the two organisms present differences, in terms of k-mer contributions to chromatin organization. Indeed, in yeast, the number of k-mers characterizing nucleosome enriched regions is larger than the number of k-mers characterizing nucleosome depleated regions, both *in vitro* and *in vivo*. In worm, it is the viceversa, and the difference between the percentages of the two types of k-mers is much more evident, especially *in vitro* (see also Section 2 of the Supplementary Material for further details and discussion, such as the relationship between the number of k-mers in the dictionaries and the length of nucleosome maps).

A second important difference between the two organisms refers to the size of the epigenomic dictionaries between *in vitro* and *in vivo* for the same organism. In particular, while the size of DY_{VV} is 95% of DY_{VT}, the size of DW_{VV} is 42% of DW_{VT}. This shows that, in both organisms, the histone k-mer preferences *in vitro* has more variety than *in vivo*. This is in agreement with previous studies showing that sequence specificity has a stronger role *in vitro* than *in vivo* (Kaplan *et al.*, 2009; Zhang *et al.*, 2011). However, this study provides novel quantitative information on this aspect, highlighting that such a variety is much more pronounced in worm than in yeast.

	Number of k-mers	**Percentage of k-mers $-$**	**Percentage of k-mers $+$**
$\mathbf{DY_{VT}}$	2,363	0.38	0.62
$\mathbf{DY_{VV}}$	2,249	0.49	0.51
$\mathbf{DY_{VT} \cap DY_{VV}}$	1,368	0.42	0.58
$\mathbf{DW_{VT}}$	186,281	0.93	0.07
$\mathbf{DW_{VV}}$	78,842	0.75	0.25
$\mathbf{DW_{VT} \cap DW_{VV}}$	57,123	0.98	0.02

Table 1. Statistics on the epigenomic dictionaries for yeast and worm.

3.1.2 Analysis of k-mers distribution in the epigenomic dictionaries In this section we present an analysis of how the k-mers are distributed in the epigenomic dictionaries with respect to their weight, i.e., the z-score of the difference between the empirical probability distributions defined in Section 2.1. To this aim, for each epigenomic dictionary, we plot the z-score values as follows. On the x axis, k-mers are represented sorted first with respect to their corresponding z-score value, taken in nonincreasing order, and then lexicographically in case of ties. The y axis reports the z-score values. The values corresponding to k-mers of different sign are plotted with different colors. The plot for DY_{VT}, DY_{VV}, and DW_{VT}, DW_{VV} are shown in Figures 1 and 2 of the Supplementary Material, respectively.

A first observation is that the curves plotting the z-score values have the same "shape", *in vitro* and *in vivo* and for both organisms. This suggests that the general k-mer compositional structure of the sequences involved in nucleosome enrichment or depletion is not disrupted by the action of "external factors", such as trascription factors and chromatin remodellers, in both yeast and worm.

Looking at their shape, all curves present a drastic "slope change", corresponding to a specific value $\widehat{z}$ such that all the z-score values smaller than $\widehat{z}$ settle around zero, whereas the z-score values larger than $\widehat{z}$ form a curve with a very steep increase on the y axis. Therefore, *in vitro* and *in vivo* and for both organisms, k-mers significantly involved in nucleosome enrichment/depletion naturally partition themselves into two classes: with very *high* (i.e., larger than $\widehat{z}$) and very *low* (i.e., smaller than $\widehat{z}$) z-score value. An empirical estimate of the values of $\widehat{z}$ are reported in Table 3 of the Supplementary Material. It is worth to point out that k-mers with very high z-score value have also high frequency of occurrence in the considered datasets. On the contrary, k-mers with low z-score have usually very few occurrences (although this was not obvious). Therefore, we refer to k-mers in the first class as *frequent*, and to those in the second class as *rare*. It is evident from the plots that only a few k-mers are frequent (their percentage is reported in Table 3 of the Supplementary Material). Further considerations on the k-mer z-score distribution are provided in Section 2 of the Supplementary Material, where it is shown that, although the shape of curves seem to suggest a power-law distribution, this is not the case.

Despite the analogy in the curves, a closer look at the k-mers sign reveals a surprising and important finding. Indeed, in yeast, all frequent k-mers characterize nucleosome depleted regions, whereas in worm they characterize nucleosome enriched regions. The next section provides further insights regarding these k-mers.

Finally, we have computed the number of frequent k-mers which are conserved between *in vitro* and *in vivo*. The result is that, for both organisms, the percentage of frequent k-mers that are conserved is higher than the percentage of rare k-mers that are conserved. Indeed, 94% of the frequent k-mers are conserved in yeast, 78% in worm, against 60% in yeast and 72% in worm for rare k-mers, respectively. Therefore, the results suggest that the main differences between *in vitro* and *in vivo*, with reference to the k-mer compositional structure of the sequences involved in nucleosome enrichment/depletion, are due to a different arrangement of the rare k-mers in the two cases, for both organisms. This is more evident in yeast than in worm.

3.2 Analysis of the homopolymeric tracts dictionaries

	poly(dA:dT)		poly(dC:dG)	
	−	+	−	+
$\mathbf{DY_{VT}}$	0.10	0	0.21	0.20
$\mathbf{DY_{VV}}$	0.07	0	0.21	0.22
Conserved	0.80	0	0.70	0.70
$\mathbf{DW_{VT}}$	0	0.05	0.06	0.03
$\mathbf{DW_{VV}}$	0	0.01	0.06	0.20
Conserved	0	0.92	1	0.38

Table 2. Percentage of k-mers in the homopolymer tracts dictionaries, distinguished for the enriched $(+)$ and depleted $(-)$ cases. For each organism, the percentage of such k-mers which are conserved between *in vitro* and *in vivo* is also shown.

Table 2 shows, for each organism, the percentage of k-mers in the epigenomic dictionaries which are homopolymer DNA tracts (the same percentages are distinguished for $k = 1, \ldots, 9$ in Table 4 of the Supplementary Material). The third and sixth rows of Table 2 report the percentage of such k-mers which are present both *in vitro* and *in vivo*. The results highlight that poly(dA:dT) have a well defined role in determining histone sequence preferences. Indeed, as opposed to poly(dC:dG), poly(dA:dT) tracts present in each of the two organisms the same preferences with respect to nucleosome enrichment/depletion, and are well conserved between *in vitro* and *in vivo*. However, it is quite remarkable that *S. cerevisiae* and *C. elegans* predispose themselves differently for nucleosome occupancy with respect to poly(dA:dT) tracts. Indeed, poly(dA:dT) tracts characterize depletion in yeast, and enrichment in worm. To the best of our knowledge, this difference between yeast and worm has not been reported previously in the Literature.

Our findings support the emerging view, initially proposed by Lorch *et al.* (2014) and in part experimentally verified by Krietenstein *et al.* (2016), according to which poly(dA:dT) tracts are involved in active remodelling mechanisms leading to nucleosome-free regions formation. Prior to those studies, the association of poly(dA:dT) tracts to nucleosome-free regions was accredited to their stiffness (Segal and Widom, 2009; Struhl and Segal, 2013): it was argued that their presence at promoters and transcription termination sites may enforce nucleosome exclusion or instability in both yeast and worm (Radman-Livaja and Rando, 2010).

In Table 5 of the Supplementary Material the percentage of homopolymer DNA tracts that are frequent k-mers is reported, showing that all poly(dA:dT) tracts are frequent k-mers in yeast and most of them are frequent k-mers in worm. Therefore, they are among the "big players" in both organisms

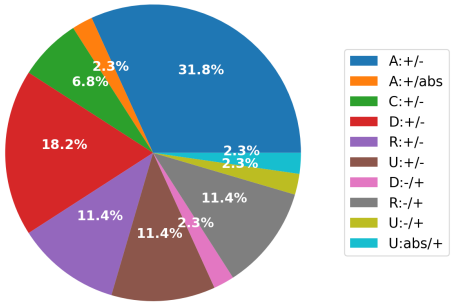

Fig. 1. Fraction of proteins for which the context of k-mer characterization changed from *in vitro* to *in vivo* for yeast. In particular, $'-'$ represents a context of depletion, $'+'$ a context of enrichment, 'abs' means that the protein is not present. Finally, A = Activators, C = Remodelers, D = Dual, R = Repressors, U = Unknown.

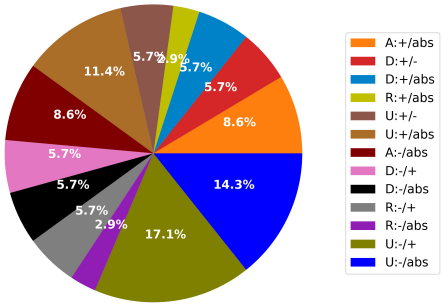

Fig. 2. Fraction of proteins for which the context of k-mer characterization changed from *in vitro* to *in vivo* for worm. The legend is analogous to that of Figure 1.

for the determination of nucleosome occupancy. Table 6 of the Supplementary Material shows instead the percentage of frequent k-mers that are homopolymer DNA tracts. In particular, most of the frequent k-mers are poly(dA:dT) tracts in yeast, while in worm about the 40% of frequent k-mers are not poly(dA:dT) tracts. Together with the lower degree of conservation of frequent k-mers between *in vitro* and *in vivo* observed in worm with respect to yeast (see the previous section), this latter fact highlights a more diversified landscape for k-mer sequence organization in worm than in yeast, with reference to histone preferences.

3.3 Analysis of the context dictionaries

Table 7 of the Supplementary Material shows that the conservation of proteins stored in the context dictionaries between *in vitro* and *in vivo* is total for yeast, in the percentage of 83% for worm.

An additional level of detail is provided by Figures 1 and 2, where the fraction of proteins for which the context has changed between *in vitro* and *in vivo* is illustrated for yeast and worm, respectively (the corresponding details are provided in Table 8 of the Supplementary Material). In particular, the context of proteins has changed in the percentage of 38% in yeast and 29% in worm. In yeast, most of the changes are in the direction from enrichment *in vitro* to depletion *in vivo*. The most common function among the proteins which show this change is that of activator. In worm, the changes from a specific context *in vitro* to the disappeareance of that protein *in vivo*, is more common than in yeast. The change from a context of depletion to a context of enrichment is slightly more common than the viceversa, although the function of the corresponding proteins is mostly unknown.

Selection has been finally applied to the context dictionaries, in order to check if they contain any poly(dA:dT) tracts. Results are shown in Tables 9 and 10 of the Supplementary Material, showing that the two organisms predispose themself differently also to the competition for preferred binding sequences. In particular, in yeast, only the 3.5% of the proteins show affinity with some poly(dA:dT) tracts, and they are all conserved between *in vitro* and *in vivo*. Moreover, their context is depletion and the most common function among them is that of activator (see Table 9 in the Supplementary Material). Therefore, we can conclude that a certain level of flexibility in favouring external factors which compete with histones for sequence occupancy (e.g., trancription factors) may be imputated to "non-poly(dA:dT)" tracts, i.e., rare k-mers (in agreement with the analysis in Section 3.2), in yeast. The scenario is more diversified in worm. Indeed, 25% of the DNA proteins show high affinity with some poly(dA:dT) tracts. Moreover, from Table 10 in the Supplementary Material, it is evident that the corresponding proteins present various functions and are characterized by different contexts, not always conserved between *in vitro* and *in vivo*.

4 CONCLUDING REMARKS

We present the first linguistic comparison between *in vitro* and *in vivo* nucleosome maps of two model organisms, based on a unified framework for the construction of compact k-mer dictionaries that allow to integrate and then analyze data coming from different sources. Moreover, we provide the first comparative analysis of *S. cerevisiae* and *C. elegans*, in terms of the sequence composition of their nucleosome maps.

Our main findings are that there is a strong sequence conservation between *in vitro* and *in vivo* in both yeast and worm and, although the two organisms present, at a high level, a similar "structure", important differences result from a deeper analysis. In particular, it is possible to identify a small "core" of frequent k-mers, and a large family of rare k-mers, for both of them. However, in yeast the core of frequent k-mers mainly includes poly(dA:dT) tracts, characterizes nucleosome depletion and is not involved in the main differences between *in vitro* and *in vivo*. Such differences may be imputable instead to rare k-mers, which seem also to be involved in the competition for sequence occupancy between histones and other factors (e.g., trascription factors and chromatin remodellers). In worm, the scenario is much more diversified: the overlap between frequent k-mers and poly(dA:dT) tracts is large but not complete, and, what is most surprising, such k-mers characterize nucleosome enrichment. This latter finding is in line with, and support even more, recent studies showing

that poly(dA:dT) tracts play a key role as *hallmarks signaling where* an active mechanism for the ATP-dependent removal of nucleosomes must be activated, as opposed to the passive role delegating them either to prevent or to form unstable nucleosomes because of their stiffness (Lorch *et al.*, 2014; Krietenstein *et al.*, 2016).

The differences presented here between yeast and worm agree with recent studies by Tompitak *et al.* (2017), showing that unicellular and multicellular organisms have opposite tendencies in nucleosome positioning sequence preferences. However, there is an important difference between their and our study. Their study is based on *in silico* occupancy maps, i.e., maps that have been obtained via a mathematical model. In this study, such a difference between unicellular and multicellular organisms is obtained via experimentally determined maps, and it applies to both intrinsic (i.e., *in vitro*) and *in vivo* organization.

Our results open new challenges, such as the identification of other basic families of k-mers within the ones discussed here, playing specific roles in chromatin organization, and possibly changing across different organisms. Other aspects which deserve further investigation, based on the framework proposed here, are: (1) analyze the possible predictive power of k-mer features (Awazu, 2017; Lo Bosco, 2016); (2) extend the approach in order to consider also other types of subwords, such as maximal motifs (Furfaro *et al.*, 2017; Pizzi *et al.*, 2018; Rombo, 2012) and/or k-mers with wildcards (Pizzi, 2016); (3) provide efficient tools for dictionary construction in the distributed via efficient k-mer statistics computation (Petrillo *et al.*, 2018).

ACKNOWLEDGEMENTS

Funding: The research by R. G. and S. E. R. has been partially supported by the InDAM Projects GNCS 2017 and GNCS 2018 of which they are members.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, New York City, NY, USA.

Awazu, A. (2017). Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition. *Bioinformatics*, **33**(1), 42–48.

Berger, M. and Bulyk, M. (2006). Protein binding microarrays (PBMs) for rapid, high- throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol*, **338**, 245–260.

Charoensawan, V., Janga, S. C., Bulyk, M. L., Babu, M. M., and Teichmann, S. A. (2012). DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. *Molecular Cell.*, **47**(2), 943–944.

Consortium, T. U. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acid Research*, **45**(D1), D158–D169.

Furfaro, A., Groccia, M. C., and Rombo, S. E. (2017). 2D motif basis applied to the classification of digital images. *Computer Journal*, **60**(7), 1096–1109.

Fuxman Bass, J. I. *et al.* (2016). A gene-centered C. elegans protein–DNA interaction network provides a framework for functional predictions. *Molecular Systems Biology*, **12**.

Giancarlo, R., Rombo, S. E., and Utro, F. (2014). Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. *Briefings in Bioinformatics*, **15**, 390–406.

Giancarlo, R., Rombo, S. E., and Utro, F. (2015). Epigenomic k-mer dictionaries: shedding light on how sequence composition influences in vivo nucleosome positioning. *Bioinformatics*, **31**, 2939–2946.

Giancarlo, R., Rombo, S. E., and Utro, F. (2018). DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes. *Theoretical Computer Science*.

Grabherr, M. G., Haas, B. J., Yassour, M., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644.

Hansen, J. C. (2012). Human mitotic chromosome structure: what happened to the 30-nm fibre? *EMBO Journal*, **31**(7), 1621–1623.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C., Pugh, B., and Korber, P. (2016). Genomic nucleosome organization reconstituted with pure proteins. *Cell*, **167**, 709–721.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, **128**, 707–719.

Lo Bosco, G. (2016). Alignment-free dissimilarities for nucleosome classification. In *Proc. of CIBB, LNCS*, volume 9874, pages 114–128.

Locke, G., Haberman, D., Johnson, S. M., and Morozov, A. V. (2013). Global remodeling of nucleosome positions in C. elegans. *BMC Genomics*, **14**, 284.

Lorch, Y., Maier-Davis, B., and Kornberg, R. D. (2014). Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes & Development*, **28**, 2492–2497.

Petrillo, U. F., Roscigno, G., Cattaneo, G., and Giancarlo, R. (2018). Informational and linguistic analysis of large genomic sequence collections via efficient hadoop cluster algorithms. *Bioinformatics*, **34**(11), 1826–1833.

Pizzi, C. (2016). MissMax: alignment-free sequence comparison with mismatches through filtering and heuristics. *Al. for Mol. Biol.*, **11**, 6.

Pizzi, C., Ornamenti, M., Spangaro, S., Rombo, S. E., and Parida, L. (2018). Efficient algorithms for sequence analysis with entropic profiles. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **15**(1), 117–128.

Radman-Livaja, M. and Rando, O. (2010). Nucleosome positioning: how is it established, and why does it matter? *Develop. Biol.*, **339**(2), 258–266.

Razin, S. V. and Gavrilov, A. A. (2014). Chromatin without the 30-nm fiber: Constrained disorder instead of hierarchical folding. *Epigen.*, **9**(5), 653–657.

Ricci, M., Manzo, C., Garca-Parajo, M. F., Lakadamyali, M., and Cosma, M. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes *in vivo*. *Cell*, **160**, 1145– 1158.

Robinson, P. J. J., Fairall, L., Huynh, V. A. T., and Rhodes, D. (2006). EM measurements define the dimensions of the 30-nm chromatin fiber: Evidence for a compact, interdigitated structure. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(17), 6506–6511.

Rombo, S. E. (2012). Extracting string motif bases for quorum higher than two. *Theor. Comput. Sci.*, **460**, 94–103.

Segal, E. and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current Opinion in Struct. Biology*, **19**, 65 – 71.

Struhl, K. and Segal, E. (2013). Determinants of nucleosome positioning. *Nat Struct Mol Biol*, **20**, 267–273.

Tompitak, M., Vaillant, C., and Schiessel, H. (2017). Genomes of multicellular organisms have evolved to attract nucleosomes to promoter regions. *Biophysical Journal*, **112**(3), 505–511.

Tremethick, D. J. (2007). Higher-order structures of chromatin: The elusive 30 nm fiber. *Cell*, **128**(4), 651–654.

Utro, F., Di Benedetto, V., Corona, D. F., and Giancarlo, R. (2016). The intrinsic combinatorial organization and information theoretic content of a sequence are correlated to the DNA encoded nucleosome organization of eukaryotic genomes. *Bioinformatics*, **32**, 835–842.

Zhang, Z., Wippo, C., Wal, M., Ward, E., Korber, P., and Pugh, B. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, **332**, 977980.

Zhbannikov, I. Y., Hunter, S. S., Settles, M. L., and Foster, J. A. (2013). SlopMap: a software application tool for quick and flexible identification of similar sequences using exact k-mer matching. *Journal of Data Mining in Genomics and Proteomics*, **4**(3), 10.4172/2153–0602.1000133.