Bioinformatics, 31(18), 2015, 2939–2946
doi: 10.1093/bioinformatics/btv295
Advance Access Publication Date: 25 May 2015
Original Paper

Sequence analysis

Epigenomic *k*-mer dictionaries: shedding light on how sequence composition influences *in vivo* nucleosome positioning

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

[1]Dipartimento di Matematica ed Informatica, Università degli Studi di Palermo, 90123 Palermo, Italy and [2]Computational Biology Center, IBM T. J. Watson Research, NY 10598, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, all the three authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on October 30, 2014; revised on March 19, 2015; accepted on May 4, 2015

Abstract

Motivation: Information-theoretic and compositional analysis of biological sequences, in terms of *k*-mer dictionaries, has a well established role in genomic and proteomic studies. Much less so in epigenomics, although the role of *k*-mers in chromatin organization and nucleosome positioning is particularly relevant. Fundamental questions concerning the informational content and compositional structure of nucleosome favouring and disfavoring sequences with respect to their basic building blocks still remain open.
Results: We present the first analysis on the role of *k*-mers in the composition of nucleosome enriched and depleted genomic regions (NER and NDR for short) that is: (i) exhaustive *and* within the bounds dictated by the information-theoretic content of the sample sets we use and (ii) informative for comparative epigenomics. We analize four different organisms and we propose a paradigmatic formalization of *k*-mer dictionaries, providing two different and complementary views of the *k*-mers involved in NER and NDR. The first extends well known studies in this area, its comparative nature being its major merit. The second, very novel, brings to light the rich variety of *k*-mers involved in influencing nucleosome positioning, for which an initial classification in terms of clusters is also provided. Although such a classification offers many insights, the following deserves to be singled-out: short poly(dA:dT) tracts are reported in the literature as fundamental for nucleosome depletion, however a global quantitative look reveals that their role is much less prominent than one would expect based on previous studies.
Availability and implementation: Dictionaries, clusters and Supplementary Material are available online at http://math.unipa.it/rombo/epigenomics/.
Contact: simona.rombo@unipa.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Studies investigating the information-theoretic content and composition, in particular in terms of *k*-mers, of sequences are pervasive in computational biology (Giancarlo *et al.*, 2009, 2012, 2014). However, although it is well established that DNA sequence has a role in epigenomics (Whitaker, 2014; Yuan, 2012), there are very few studies that have systematically applied compositional and linguistic techniques for the identification of sequence features associated with epigenomic functions. They are reviewed by Pinello *et al.* (2014). To the best of our knowledge, no information-theoretic study is present in epigenomics.

We focus on the following particular research area in epigenomics: the identification of mechanisms accounting for nucleosome organization and positioning in chromatin. It was initiated more than 30 years ago by Kornberg (1981) and, thanks to advances in microarray and sequencing technologies, it has prospered in the past few years. An extensive overview of its many aspects is provided in (Jiang and Pugh, 2010; Minary and Levitt, 2014; Radman-Livaja and Rando, 2009; Segal and Widom, 2009b; Struhl and Segal, 2013). Particularly relevant for this paper are the findings outlined next.

The chromatin organization of eukaryotic genomes has been successfully argued that it is DNA-encoded (Kaplan *et al.*, 2008). Therefore, it seems natural to complement such an argument with an information-theoretic study aiming at establishing the similarity and differences, in terms of their information-theoretic content, between NER and NDRs on a genomic scale. Such a study is not available. It is also a well established fact that sequence motifs and regularities influence nucleosome positioning. The ones that have been identified are: (i) the 10 bp periodicity of the dinucleotides AA/TT/TA that oscillate in phase with each other and out of phase with a similar perodicity of the GC denuclotides (Segal *et al.*, 2006); (ii) poly(dA:dT) tracts (Segal and Widom, 2009a), i.e. stretches of A's or T's and (iii) the G + C content of a genomic region, with its A + T content also playing some role (Peckham *et al.*, 2007; Tillo and Hughes, 2009). In terms of specific *k*-mers, studies about their role in favouring nucleosome positioning have only identified a handful of them, e.g. (Peckham *et al.*, 2007; Tillo and Hughes, 2009) and no systematic study seems to be available.

The above state of the art clearly indicates that the literature has not addressed the following foundational issues:

1. An information-theoretic characterization of the NER and NDR, highlighting similarities and differences.
2. A structural characterization of the 'world of *k*-mers' in terms of their involvement in the composition of genomic sequences rich or depleted of nucleosomes, with the possible identification of groups of similar *k*-mers that play a role in that composition.

We provide contributions to both issues. In particular, we present the first information-theoretic analysis of NER and NDR. Among the other useful insights, it establishes that the information content of NER is surprisingly very close to that of NDR, yet such a difference is statistically very significant. As for the second issue, we provide a computational and statistical methodology that is used to build *epigenomic dictionaries*, for the case of nucleosome positioning *in vivo*. That is, catalogues of *k*-mers, each having a statistical score assessing to which extend it favours or inhibits nucleosome formation. Based on a sound information-theoretic argument, *k*-mers of at most 10 basis are considered in this study. Interestingly, *k*-mers that are known to favour or inhibit nucleosome positioning from previous studies are correctly classified in our dictionaries, and a rich, never highlighted before, variety of *k*-mers involved in nucleosome positioning comes to light for the first time.

A final introductory remark is in order. The only previous study, of which this one can be seen as its natural continuation, is the one by Peckham *et al.* (2007), where *in vivo* data have been used. We adhere to that choice here, pointing out the need for a genome-wide analysis comparing *k*-mer preferences for nucleosome formation *in vitro* and *in vivo*. The very discriminative and methodologically sound techniques given here are a relevant technical step forward for its successful realization, whose outcome would add another important tile to the puzzle of 'sequences and nucleosomes'.

2 Methods

The statistical and computational methodologies needed to address points (1) and (2) of the Section 1 are summarized here. In particular, Section 2.1 highlights some basic steps, either identical or variants of well known ones in sequence analysis. Section 2.2 is dedicated to the definition of weighted *k*-mer dictionaries, together with weighting schemes, i.e. procedures to assign weights to *k*-mers. Such a data structure is pragmatically used in genomic research, but it has not been formalized. Our experiments bring to light the novelty that such a pragmatism *hides* a powerful methodological paradigm: properly modulated via an associated weighting schema, a weighted dictionary can provide information about the same data on different and complementary scales. Such a somewhat surprising modulation ability, via two apparently very similar weighting schemes, is demonstrated in Section 3.2 by comparing the results obtained via two weighting schemes, one already used in nucleosome positioning studies (Peckham *et al.*, 2007), the other novel.

2.1 *K*-mer probabilities from counts: estimation, conservation and differences in information content

Let Σ be an alphabet, fix an integer $k \geq 1$ and let $\mathbf{P}_k$ be a probability distribution establishing how probable is the extraction of any given k-mer in Σ^k, to form a set of strings. Using a sample set D for the extraction process, an estimate $\widehat{\mathbf{P}}_k$ of $\mathbf{P}_k$ needs to be computed. This task is nearly standard in sequence analysis (Durbin *et al.*, 1998). Intuitively, one would like to use the *k*-mer frequency counts obtained from the sample set. Indeed, the corresponding empirical probability distribution is a Maximum Likelihood estimate of $\mathbf{P}_k$. However, when the sample size is too small compared with Σ^k, there may be rare or missing *k*-mers in the sample which, having zero or close to zero frequency counts, give rise to *k*-mer probabilities with zero value in the estimate. In those cases, in order not to rely on a small sample size for the estimation, one resorts to the introduction of suitably chosen pseudo counts: they are added to the frequencies of the *k*-mers. They are obtained by selecting an appropriate *prior* distribution encoding prior knowledge about $\mathbf{P}_k$. While the book by Durbin *et al.* (1998) offers a good introduction to this topic, for the convenience of the reader and to keep the paper self-contained, some basic technical details are given in Section 1 of the Supplementary Material. For this study, five of the most established *priors* in the literature have been used: namely Maximum Likelihood Estimate, Uniform-Bayes-Laplace, Jeffreys (1946), Perks (1947) and Trybula (1947) (see Table 1 of the Supplementary Material).

Once fixed the sample set D, it is to be expected that, as k grows, the number of *k*-mers rare or absent in D increases. That results in an estimation of $\mathbf{P}_k$ where prior knowledge (via the pseudo-counts) becomes more and more relevant. To avoid such a problematic estimation, it is natural to ask up to which maximum value $k_{\max}$ a good estimate can be granted. It is worth of mention that such a question is usually dealt with heuristically. Here a principled and quantifiable choice is made: the value of $k_{\max}$ to be selected must guarantee an estimate of $\mathbf{P_k}$ that accurately represents the information-theoretic content of D. That is, the sample size should be 'big enough' to allow for a good estimation of the entropy of the source generating the sample. The value of $k_{\max}$ here is computed according to a procedure recommended in (Dudok de Wit, 1999) for entropy estimation. Again, for the interested reader and to keep the presentation self-contained, technical details are given in Section 1 of the Supplementary Material. The only thing that is needed at this stage is that the procedure takes as input the dataset and a fixed *a priori*

Table 1. Percentage of *k*-mers stored in the base dictionaries (columns 2 and 3), with respect to the total number of *k*-mers considered in this study

	HT-WD		AUC-WD	
ORGANISM	+	–	+	–
yeast	3.49	1.197	0.01	0.013
human	4.144	0.543	0.008	0.008
fly	3.128	1.56	0.005	0.006
worm	4.576	0.111	0.004	0.01

The abbreviations HT-WD and AUC-WD stand for Hypothesis Test and Binary Classification weighted dictionaries, respectively (also in the next tables).

threshold value ϵ in [0,1], this latter quantifying the tolerable percent difference between the entropy one could estimate from the data and the true entropy of the source emitting the data. The closer ϵ is to zero, the better the estimate is. The output of the procedure is $k_{\max}$, for the given ϵ and sample set.

Finally, let **Q** and **P** be two *k*-mer probability distributions. To establish how close **Q** and **P** are in terms of their 'information content' the following can be used: (i) the Hellinger distance D_{HL} (Deza and Deza, 2006), that assumes values in the interval [0,1]; (ii) a dissimilarity measure S_{KL}, based on the Kullback-Leibler divergence (Cover and Thomas, 1991). They are both measures of difference in information content in probability distributions (Csizár, 1967) and their formal definition is provided in Section 1 of the Supplementary Material. For completeness, we mention that those as well as analogous measures have been used in related studies in epigenomics (e.g. Pinello *et al.*, 2011, 2014).

2.2 Weighted *k*-mer dictionaries: the special case of nucleosome positioning

Let $\alpha \in (0,1)$ be a real value and fix an integer $k \in [1, k_{\max}]$. Let $\mathcal{D}_{k,\alpha}$ denote a set of triplets $\langle x, w, s\rangle$ such that: x is a *k*-mer, w is a real value such that $\alpha \leq w \leq 1$ and s is a symbol from the binary alphabet $\{+,-\}$. Intuitively, a value of $s = \pm$means that x is a 'characteristic/significant feature' of NER/NDR. In the reminder of this paper, to refer to such an intuition, we use the shorthand '*k*-mer favouring/disfavouring nucleosome positioning'. Again intuitively, the entire triple states that x favours/disfavours nucleosome formation with a 'confidence level' w at least equal to the given threshold α. The set $\mathcal{D}_{k_{\max},\alpha} = \cup_{k=1}^{k_{\max}} \mathcal{D}_{.k,\alpha}$ is a *weighted k-mer dictionary*. When no ambiguity arises, it will be referred to simply as *dictionary*.

The 'semantic' of a dictionary is given by a weighting scheme, which is a procedure that assigns weights to the *k*-mers in a dictionary suitably designed to assess via data analysis the level of involvement of *k*-mers in nucleosome positioning.

Given two dictionaries, $\mathcal{D}_{\hat{k}_{\max},\hat{\alpha}}$ and $\mathcal{D}_{k'_{\max},\alpha'}$, let $\alpha = \min(\hat{\alpha}, \alpha')$ and $k = \min(\hat{k}_{\max}, k'_{\max})$. The `join` of those two is a new dictionary $\mathcal{D}_{k_{\max},\alpha}$ obtained by taking all *k*-mers common to both, with the same sign and a confidence level at least α.

For a given organism, the weighted dictionary that can be built directly from genome-wide nucleosome positioning maps (referred to simply as *maps*, when no ambiguity arises) has the special role of a `base dictionary`. It can also be obtained by joining several dictionaries, each obtained with the use of a distinct map.

We point out that the definition given above can be easily put in general terms. The details are left to the reader.

2.2.1 The choice of a weighting scheme for nucleosome positioning

Intuition suggests that, given a *k*-mer x favouring nucleosome enrichment (to fix ideas), one of the following, non-mutually exclusive, things should happen: its frequency should be (a) able to classify well $F = E \cup D$ into E and D when the frequency of x in $f \in F$ is used as a classification score; (b) 'significantly' different in NER and NDR, i.e. such a difference in frequency is not due to chance.

Intuitively (i) can be formalized via Binary Classification in Machine Learning while (ii) via Hypothesis Test in Statistics and, for the convenience of the reader, they are both reported next. It is worth to mention that (i) has been used by Peckham *et al.* (2007) in extracting sequence nucleosome positioning signals in *S.cerevisiae* while, to the best of our knowledge, the use of (ii) to 'rank' *k*-mers in terms of their favouring/disfavouring nucleosome positioning preferences, is novel.

For completeness, it is worth mentioning that the identification of an appropriate weighting scheme is strongly related to *feature selection* in machine learning (Guyon and Elisseeff, 2003), although neither of the two schemes outlined next can be regarded as feature selection technique. This remark poses the problem of investigating feature selection techniques in the context of this paper.

A weighting scheme based on Binary Classification

Fix a *k*-mer x. Each sequence in F is given a score equal to the frequency of occurrence of x in it, normalized by its length. Those scores are then used to evaluate how well they classify E and D, via ROC analysis (Fawcett, 2006). To this end, the analysis is first performed by assigning class label 0 to sequences in E and then class label 1. Notice that the assignment of a class label to sequences in E determines the assignment of the corresponding class label to sequences in D. The maximum of the two corresponding AUCs is assigned as a confidence level to x. The symbol s is set to '+' if the AUC with class label 1 assigned to E is higher than the AUC with class label 0 assigned to E, and to '–' otherwise. The threshold α is a real number in $[0.5, 1)$ and corresponds to the minimum AUC that a *k*-mer must obtain to be included in the dictionary.

A weighting scheme based on Hypothesis Test

Let **Q** and **P** be the *k*-mer empirical probability distributions associated to the sample sets E and D, respectively. For each x, let $d_x = |p_x - q_x|$. Such a difference is normalized via the z-score z_x (see, e.g. Triola, 2012 for the definition of z-score and its uses in data normalization). To establish the statistical significance of z_x, a Hypothesis Test can be performed via a Montecarlo simulation. The interested reader can find details in (Giancarlo *et al.*, 2008; Giancarlo and Utro, 2012; Gordon, 1996). The Null Hypothesis that the value of z_x is due to chance is formalized by the way in which the artificial datasets E' and D' (corresponding to E and D, respectively) are generated in each step of the simulation. In particular, the set $F = E \cup D$ is first shuffled a certain number of times (1000 times, in this case) and then splitted in the two sets E' and D', with $|E'| = |E|$ and $|D'| = |D|$. The symbol s is set to '+' if $p_x > q_x$, and '–' otherwise. The threshold α is set to the significance level used in the test to reject the null hypothesis.

3 Results and discussion

We present here our findings in relation to the problems posed in the Introduction. In particular, our analysis is based on the maps of

four different organisms: *S.cerevisiae* (yeast), *D.melanogaster* (fly), *H.sapiens* (human) and *C.elegans* (worm).

3.1 Distinguishability of NER and NDR based on their information content

We adopt the methodology in Kaplan *et al.* (2008) (for more details, see Section 2 of the Supplementary Material): They investigate the intrinsic organization of eukaryotic genomes, remarkably showing that *in vitro* maps can be used to reliably distinguish NER from NDR produced from *in vivo* maps. Here *in vivo* NER and NDR are studied, to quantify how distinguishable they are in terms of their intrinsic information-theoretic content. We used the normalized map for *in vivo S.cerevisiae* produced by those authors and the adjusted occupancy map for *C.elegans* by Valouev *et al.* (2008) (as in Kaplan *et al.*, 2008, only chromosome 2). NER and NDR sets are extracted from those two maps.

For both *S.cerevisiae* and *C.elegans*, the sample sets of relevance here are the NER, NDR and the genome sequence underlying the map, respectively. For each of the mentioned sample sets, its k-mer probability distribution has been estimated using each of the pseudocounts mentioned in Section 2.1 and given explicitly in Table 1 of the Supplementary Material. The choice of one class of pseudocounts, rather than another, does not affect the conclusions that can be drawn from the corresponding experiments. Therefore, the presentation of our results is limited only to probability distributions obtained from Jeffreys pseudocounts. The other results are available in Section 3 of the Supplementary Material, Figures S2–S5.

To allow for a comparative analysis with the same range of k's on all sample sets used in this paper, the estimation of $k_{\max}$ has been performed, via the procedure mentioned in Section 2.1, including the sample sets in Section 3.2. Setting $\epsilon = 2.1\%$, i.e. a very conservative estimation, the result is $k_{\max} = 10$. Therefore, k-mers long at most 10 have been considered in all the experiments reported in this paper. The details on how such a value has been obtained, and a discussion about the relation between the various values of k and ϵ, are in Section 3 of the Supplementary Material.

Via D_{HL} and S_{KL}, the following comparisons have been made, each involving comparisons between probability distributions computed with the Jeffreys pseudo-counts and using the regions we list next as sample sets: (i) each of NER and NDR versus the corresponding genome; (ii) the NER versus the NDR of each organism; (iii) the NER versus NER and the NDR versus NDR of the two different organisms. While comparisons (i) and (ii) are quite natural and do not need explanation, (iii) is performed because, for the first time, it would provide a quantitative assessment of the extend of information-theoretic conservation between organisms in regard to nucleosome positioning signals in genomic DNA.

The results are reported in Figure 1. It is evident that the growth of D_{HL} is steeper than that of S_{KL} and, for both, the reported differences are small. Although the first observation may be related to the different nature of the two functions, for which no known mathematical relationship is available (and in fact they were chosen exactly for that reason: to get two independent 'readings' of the informational differences in the sample sets), the second fact is important and an assessment of the statistical significance of those differences is required. Therefore, a Hypothesis Test has been performed, where the Null Hypothesis is that those differences are due to chance and it is formalized according to the methodology explained in Section 2.2.1. With use of the Jeffreys pseudo-counts, a Monte Carlo simulation has been performed (20 steps), to determine in which case the

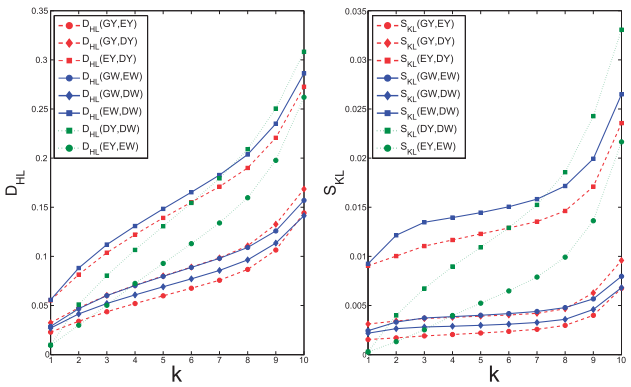

Fig. 1. The results of the computation of D_{HL} (left) and S_{KL} (right) performed on the NER, NDR and genomes of yeast and worm. They are depicted according to the legend in the figure as follows. EY, DY and GY stand for NER, NDR and genome in yeast. The abbreviations for worm have analogous meaning. Moreover, to improve the figure 'readability', the arguments of both D_{HL} and S_{KL} are the relevant sample sets instead of their Jeffreys pseudo counts

Table 2. Percentage of k-mers stored in the `join` of the base dictionaries, with respect to the size of the smallest set considered for the `join`

	HT-WD		AUC-WD	
`join`	+	−	+	−
yeast, worm	24	13	50	80
yeast, fly	19	2	50	40
yeast, human	21	2	50	60
fly, worm	16	0.9	67	63
fly, human	40	13	40	83
human, worm	14	0.9	75	75
yeast, worm, human, fly	2	0	25	40

Null Hypothesis can be rejected with a very high significance level, i.e. 1% threshold. The results are reported in Tables 2 and 3 of the Supplementary Material, where the captions specify to which distance they refer to. With a few minor exceptions at $k = 10$, which may be considered a border-line case in terms of conservation of information, all of the computed difference values are highly significant. Therefore, using two functions with no known mathematical relation between them, we get the same consistent result: small differences and statistically significant, yielding very robust results for k in [1,9] that, in turn, lead to the following remarkable conclusions:

Although the sets involved in the comparison carry essentially the same amount of information (indeed the distance values in Fig. 1 are quite low), the Hypothesis Test shows that those small differences are statistically significant. Therefore, those sets are distinguishable in information-theoretic terms. Surprisingly, such an assertion applies also to the comparison of homologous nucleosome enriched/depleted regions between the two organisms indicating that, although small, there is a significant level of differentiation in the information content of those regions between organisms.

As indicated by Figure 1, the curves associated to D_{HL} and S_{KL} for NER and NDR in the same organism are translated top-ward with respect to the case of NER/NDR versus the respective genomes. The distance values differ even by one order of magnitude for S_{KL}. That indicates a significant and higher level of differentiation in

Table 3. The *k*-mers in common to all organisms and coming from the Binary Classification dictionaries

Common k-mers from AUC-WDs							
GC	+	CAC	+	CC	+	AC	+
CAA	+	ACG	+	CAG	+	AGTA	+
GAC	+	TCAA	+	GCA	+	CATC	+
CA	+	ACC	+	G	+	CGA	+
AT	–	TAA	–	AAAT	–	ATTTA	–
AAA	–	AAAT	–	A	–	AATA	–
ATAA	–	AATT	–	AAAAT	–	AAATT	–
AAAAAA	–	AAAAAT	–	ATAAA	–	AATAT	–
GAAAA	–	TAAAAA	–	ATTTA	–	AAAATA	–
AA	–	AAAA	–	TAAA	–	AAAAA	–
TAAAA	–	C	+	ATAA	–	AAATA	–

information-theoretic terms between those regions with respect to each of them compared with the 'background' genome.

Figure 1 shows a non-decreasing trend in the value of the differences in information content, as a function of *k*. This behaviour is more accentuated for the curves corresponding to homologous nucleosome enriched/depleted regions between the two organisms. Although we found an analogous growth also in randomly generated sequences (see Fig. S6 in the Supplementary Material), it is much more prominent in genomic sequences. Therefore, this establishes the following novel fact, relevant for syntactic-linguistic studies of biological sequences. As the length of their 'building block' grows, there is a growing differentiation in terms of the information content of the genomic regions studied here, indicating a growing syntactic-linguistic difference in the organization of the basic building blocks within those sequences.

3.2 Epigenomic dictionaries

For the extraction of NER and NDR from nucleosome positioning maps, a valid alternative to the one proposed by Kaplan *et al.* (2008) is well exemplified by Valouev *et al.* (2008) in which nucleosome core and linker regions are used to investigate the existence of universal sequence features, specifically short *k*-mers, involved in nucleosome positioning. A data set of this kind has been recently set-up by Guo *et al.* (2014), for research only somewhat related to our study, but it is an excellent choice for an up-to-date collection of short nucleosome-favoring/disfavoring sequences. Those authors provide three pairs of sets of short NER and NDR associated to *C.elegans, D.melanogaster* and *H.sapiens*, respectively. Indeed, each pair is extracted from genome-wide *in vivo* nucleosome positioning maps of the corresponding organism. For this type of data set, the NER are (typically) nucleosome core sequences and the NDR are linker regions, each of length 147-bp. A summary of the procedure used to extract those data sets from the corresponding maps is reported in Section 2 of the Supplementary Material.

3.2.1 Weighted dictionaries: base and inter-organism

Epigenomic dictionaries, for the specific case of *k*-mers involved in the composition of nucleosome forming/disfavoring sequences, are obtained with the use of both the NER and NDR of Section 3.1 and the ones obtained from the data by Guo *et al.* (2014). Moreover, data derived by Peckham *et al.* (2007) for *S.cerevisiae* has also been integrated in our framework, since their supplementary file 'feature scores' is a *de facto* dictionary with a Binary Classification weighting scheme. For each organism, the procedures described in Section 2.2 have been applied with thresholds at 16% and 0.55 to obtain the corresponding weighted dictionaries, respectively. This means that only those *k*-mers whose confidence level, i.e. either p-value or AUC, is higher than or equal to the fixed thresholds are considered of relevance and therefore stored in the corresponding dictionary. Based on those, two different kinds of dictionaries have been built, as follows.

- Base dictionaries with Binary Classification and Hypothesis Test weighting schemes. For fly and human, the dictionaries mentioned earlier are taken as base. This applies also to yeast, but only for the Hypothesis Test case. As for its base Binary Classification dictionary, it is the `join` of the dictionaries coming from the NER and NDR in Section 3.1 and the mentioned file by Peckham *et al.* (2007). As for the worm, for both types of dictionary, we take as base the `join` between the dictionary coming from the NER and NDR in Section 3.1 and that coming from the NER and NDR obtained from the data in (Guo *et al.*, 2014).
- Inter-organism. For each subset of at least two of the four organisms considered here, both types of dictionaries have been built to explore conservation of *k*-mer involvement in different organisms. They are the `join` of the corresponding base dictionaries.

3.2.2 Summary statistics

Table 1 shows the percentage of *k*-mers stored in the base dictionaries (columns 2 and 3), with respect to the total number, i.e. $\approx 4^{10}$, of *k*-mers that have been tested for inclusion. Table 2 (Columns 2 and 3) provides statistics shedding light on the level of conservation in *k*-mer usage for the composition of NER and NDRs. How that is measured is explained in the table caption. In all tables presented here nucleosome forming *k*-mers are indicated with a '+', while '–' denotes the nucleosome disfavouring *k*-mers. Additional statistics on the number of *k*-mers stored in the dictionaries are given in Tables S4 and S5 of the Supplementary Material.

Even with confidence thresholds only moderately selective, both techniques bring to light that a very low percentage of *k*-mers are involved in nucleosome formation/depletion. Moreover, nucleosome forming *k*-mers are much more abundant than nucleosome disfavouring ones (see Table 1 again). Among and between different organisms, there is some level of conservation on the usage of nucleosome forming or disfavoring *k*-mers. However, the set(s) in which there is agreement is rather small and quite dependent on the organisms involved in the `join`.

Those two tables also bring to light the complementarity of the two techniques used for this study. Indeed, the one based on Binary Classification is certainly much more discriminative than the one based on Hypothesis Test. Its main merit is its ability to identify in the corresponding dictionaries of each organism a small kernel of *k*-mers that are common in usage either between two organisms or among all of them.

On the other hand, the Hypothesis Test technique seems to be able to capture subtler '*k*-mer nucleosome positioning signals', since the *k*-mer percentage is larger than the one contained in the Binary Classification homologous. As it is discussed in the remainder of this study, such a complementarity allows to obtain additional sequence specific information with respect to what was already known in the literature, while opening the way to explore a variety of *k*-mer usage not available before and that accounts for organism specificity.

In the next two subsections, the dictionaries will be used to show results coming from those two complementary levels of detail.

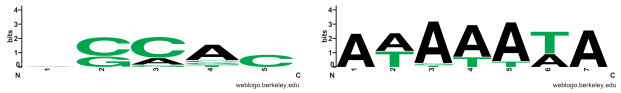

Fig. 2. Sequence logos of the alignment of *k*-mers in Table 3, distinguishing those favouring (left) from those disfavouring (right) nucleosome positioning

3.2.3 Binary classification dictionaries: a high level detail

A small kernel of *k*-mers significantly involved in the composition of NER and NDRs, common to all organisms is reported in Table 3. It has been obtained via a `join` of the base dictionaries studied in this section. As evident from the discussion that follows, that kernel provides high level details about the role of *k*-mers in nucleosome positioning, specifically addressing the somewhat neglected topic of conservation among organisms.

Figure 2 provides the sequence logos (Schneider and Stephens, 1990) obtained by aligning with CLUSTALW (Thompson *et al.*, 1994) the nucleosome favouring and disfavouring *k*-mers. As evident from the corresponding sequence logo, a poly(dA:dT) tract emerges as a common feature of nucleosome disfavouring *k*-mers. This is well known in the literature (Segal and Widom, 2009a). On the other hand, no such a definite high level common pattern emerges for nucleosome favouring *k*-mers. Those facts suggest that, while poly(dA:dT) tracts are a 'strong signature' of nucleosome depletion shared by organisms, no analogous 'signature' seems to exist for nucleosome formation, hinting that the latter is more organism specific of the former as far as *k*-mers are concerned.

At this 'very discriminative' level of detail, our main contribution, which consists of providing novel organism and sequence specificity, clearly emerges from the comparison of our findings with the ones of two existing studies, those latter being summarized next for the convenience of the reader. They are related to this part of our study because they either use the same Binary Classification approach (Peckham *et al.*, 2007) or closely related machine learning techniques (Tillo and Hughes, 2009), i.e. the Lasso feature selection method (Tibshirani, 1996) (see again Section 2.2.1 for the relation between Binary classification and feature selection in this context).

Peckham *et al.* (2007), in their study of nucleosome positioning signals in genomic DNA, sigled out 31 *k*-mers that found of relevance there. They were obtained by analyzing *S.cerevisiae* positioning maps that are quite different than the one considered here. That study leaves open how conserved the role of those relevant *k*-mers is across organisms. Our study can be used to provide an answer to that important question. Indeed, quite remarkably, the vast majority (24 out of 31) of the *k*-mers identified in the mentioned previous study are common to all organisms considered here in the base Binary Classification dictionary. Details are in Table S6 of the Supplementary Material.

On the other hand, Tillo and Hughes (2009), in their study on the construction of a simplified model for prediction of nucleosome positioning from sequences, singled out 14 sequence features deemed important for their model. They were extracted by using an *in vitro* nucleosome map of *S.cerevisiae* by Kaplan *et al.* (2008). Eleven of such features are 4-mers. We find that seven of them appear in all of our Binary Classification dictionaries. Moreover ten out of such eleven 4-mers are present in the majority of the organisms considered here. The exception is the *k*-mer found to be 'the least relevant' in the study by Tillo and Hughes (2009). It is worth of mention that only four relevant *k*-mers are common to the study by Peckham *et al.* (2007) and Tillo and Hughes (2009). Again, details are in Table S6 of the Supplementary Material.

Table 4. For each Hypothesis Test dictionary, the number of clusters (NC) obtained via DNACLUST, their maximum (MXS) and medium (MDS) sizes

	+			−		
ORGANISM	NC	MXS	MDS	NC	MXS	MDS
yeast	335	387	31	278	382	21
human	490	1988	105	225	207	21
fly	453	687	75	373	301	36
worm	753	427	47	113	46	3

In conclusion, regarding the assessment, via machine learning techniques, of *k*-mer involvement in nucleosome positioning, our study accounts for both studies just outlined on *S.cerevisiae*. Moreover, it extends them both in terms of (i) organisms and (ii) relevant *k*-mers common to all organisms.

3.2.4 Hypothesis test dictionaries: additional levels of detail

For the dictionaries studied in this section, we limit the analysis to a very discriminative significance level, i.e. 2%. The level of detail offered by this part of our study is novel.

For each of the Hypothesis Test dictionaries, the sets of *k*-mers favouring/disfavouring nucleosome positioning, with the specified significance level threshold, have been clustered separately via DNACLUST (Ghodsi *et al.*, 2011), with a sequence similarity threshold of 75%, computed via standard semi-global alignment. Table 4 reports the number of clusters so obtained, for each organism. It gives a synopsis of the variety of similar *k*-mers involved either in favouring or disfavouring nucleosome positioning, with a very high statistical significance. To the best of our knowledge, such a classification is new in the literature.

The results of the clustering process provide a valuable hierarchical access to the information contained in each Hypothesis Test dictionary: the representatives and then the clusters. Those latter, being composed of fairly similar sequences, can be used to highlight common sequence patterns for *k*-mers involved in nucleosome positioning. Significant examples of that are in Figure 3, showing the sequence logos obtained via alignment with CLUSTALW of specific clusters, corresponding to *k*-mers favouring or disfavouring nucleosome positioning. A more extensive set of analogous logos is provided in the Supplementary Material (Figs S7–S13). As opposed to the high level detail reported in Figure 2 for nucleosome favouring *k*-mers, patterns now emerge and they highlight that the G + C content as a determinant of nucleosome positioning is only a good approximation to a much more specific and complex set of sequence patterns favouring nucleosome positioning (as an example, see the first logo in Fig. 3). Likewise, the patterns that emerge for nucleosome disfavouring *k*-mers, are very much related to poly(dA:dT) tracts only in *S.cerevisiae* and partly in *C.elegans* (see the second and the fifth logos in Fig. 3). Although those types of patterns occur in some of the considered organisms (e.g. fly and human), there are patterns that substantially diverge from being poly(dA:dT) (as shown by the third and fourth logos in Fig. 3). Quite remarkably, this is in agreement with the finding that the 5-mer AAAAA is strongly associated to nucleosome depletion in the yeast and in part in the worm, but it has a much less positioning influence in both fly and human (Radman-Livaja and Rando, 2009). Section 4.1 of the Supplementary Material offers a quantitative assessment of this point (Tables S7–S8).

Both qualitatively and quantitatively, our results suggest that short poly(dA:dT) tracts are *relatively* important for nucleosome

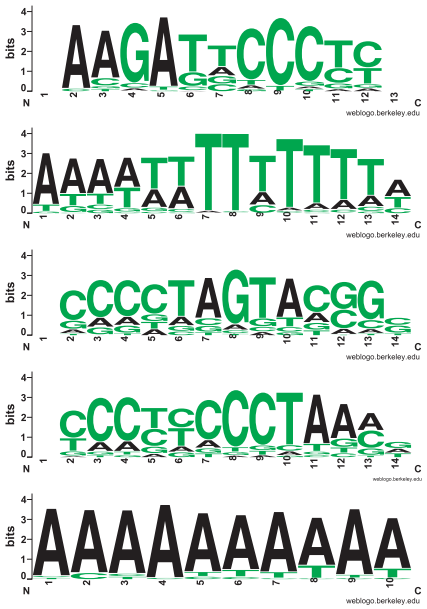

Fig. 3. Sequence logos of (from up to down): The 95 aligned sequences in Cluster 9, nucleosome favouring *k*-mers, for the worm; the 380 aligned sequences in Cluster 23, nucleosome disfavouring *k*-mers, for the yeast; the 223 aligned sequences in Cluster 12, and the 283 aligned sequences in Cluster 1, nucleosome disfavouring *k*-mers, for the fly; the 45 aligned sequences in Cluster 7, nucleosome disfavouring *k*-mers, for the worm

positioning in yeast and partly worm, in relation to other *k*-mers, but such an importance seems to vanish in fly and human.

Moreover, the presence of short poly(dA:dT) and poly(dC:dG) tracts in nucleosome forming patterns argues for the need of a better understanding of the role of DNA deformation in relation to the entire spectrum of biological processes where it plays a role, in agreement with findings in (Johnson *et al.*, 2013).

4 Conclusive remarks

The analysis reported here sheds light on how sequence composition may influence nucleosome positioning. In particular, we have found that nucleosome enriched and depleted regions are remarkably and unexpectedly close in terms of their information-theoretic content, and only small differences in their composition are responsible for their functional diversity. To understand the organization of those *k*-mers that are significant in favouring/disfavouring nucleosome positioning, we have proposed a consistent paradigm useful to distinguish them according to two different points of view: one based on Binary Classification, able to provide a more general and high level description, and the other one relying on Hypothesis Test in Statistics, able to provide more detailed information with respect to Binary Classification. Thanks to the use of those two complementary views, we have both confirmed and extended what was already known in the literature, showing that the scenario is richer in k-mers variety than one could have expected. This opens the way to several further directions of analysis. Notably among them, the application of motif discovery techniques (e.g. Rombo, 2012; Parida *et al.*, 2014) to single out possible regularities among significant *k*-mers. Moreover, special mention deserves the need for a comparative study of *k*-mer involvement in nucleosome formation coming from *in vitro* and *in vivo* maps, as already pointed out in the Introduction. That would be a natural complement and continuation of the study reported here for *in vivo* maps.

Acknowledgements

The authors are grateful to: Noam Kaplan for providing the *S.cerevisiae* maps; Davide Corona for informative discussions about chromatin organization; Chiara Romualdi for very helpful discussions on statistical methods related to this research and Simona Panni for her critical comments on a revised version of this manuscript. Finally, the authors are deeply indebted to the referees for their constructive criticism that certainly helped to make this study more accessible to the general readership of the journal.

Funding

FIRB Projects: RBNE01FSWT 'Bioinfomatica per la Genomica e la Proteomica', RBIN04BYZ7 'Algoritmi per la Scoperta ed il Ritrovamento di Patterns in Strutture Discrete, con Applicazioni alla Bioinformatica' (to R.G.). PRIN Project 20122F87B2 'Approcci composizionali per la caratterizzazione e il mining di dati omici' (to S.E.R.). All mentioned projects are financed by the Italian Ministry of Education, Universities and Research. Additional support to R.G. and S.E.R. is provided by Progetto di Ateneo (U. of Palermo) 2012-ATE-0298 'Metodi Formali ed Algoritmici per la Bioinformatica su Scala Genomica'.

Conflict of Interest: none declared.

References

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley-Interscience, New York City.

Csizár,I. (1967) Information-type measures of difference of probability distributions and indirect observation. *Studia Scient. Mathemat. Hungarica*, **2**, 229–318.

Deza,E. and Deza,M. (2006) *Dictionary of distances*. Elsevier, Amsterdam.

Dudok de Wit,T. (1999) When do finite sample effects significantly affect entropy estimates. *Eur. Phys. J.*, **11**, 513–516.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Caombridge University Press, Cambridge, UK.

Fawcett,T. (2006) An introduction to ROC analysis. *Patt. Rec. Lett.*, **27**, 861–874.

Ghodsi,M. *et al.* (2011) DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, **12**, 271.

Giancarlo,R. and Utro,F. (2012) Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis. *Theor. Comput. Sci.*, **428**, 58–79.

Giancarlo,R. *et al.* (2008) A tutorial on computational cluster analysis with applications to pattern discovery in microarray data. *Math. Comput. Sci.*, **1**, 655–672.

Giancarlo,R. *et al.* (2009) Textual data compression in computational biology: a synopsis. *Bioinformatics*, **25**, 1575–1586.

Giancarlo,R. *et al.* (2012) Textual data compression in computational biology: algorithmic techniques. *Comp. Sci. Rev.*, **6**, 1–25.

Giancarlo,R. *et al.* (2014) Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies. *Brief. Bioinf.*, **12**, 265–272.

Gordon,A. (1996) Null models in cluster validation. In: Gaul,W. and Pfeifer,D. (eds.), *From Data to Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization*. Springer Berlin Heidelberg, pp. 32–44.

Guo,S. *et al.* (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, **30**, 1522–1529.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Jeffreys,H. (1946) An invariant form for the prior probability in estimation problems. *Proc. Roc. Soc. (Lond.) A*, **186**, 453–461.

Jiang,C. and Pugh,B. (2010) Nucleosome positioning and gene regulation: advances through genomics. *Nat. Genet.*, **10**, 161–172.

Johnson,S. *et al.* (2013) Poly(dA:dT)-Rich DNAs are highly flexible in the context of DNA looping. *PLoS One,* **8**, e75799.

Kaplan,N. *et al.* (2008) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

Kornberg,R.D. (1981) The locations of nucleosomes in chromatin: specific or statistical? *Nature*, **292**, 579–580.

Minary,P. and Levitt,M. (2014) Training-free atomistic prediction of nucleosome occupancy. *Proc. Natl Acad. Sci.*, **111**, 6293–6298.

Parida,L. *et al.* (2014) Irredundant tandem motifs. *Theor. Comput. Sci.*, **525**, 89–102.

Peckham,H.E. *et al.* (2007) Nucleosome positioning signals in genomic DNA. *Genome Res.*, **17**, 1170–1177.

Perks,W. (1947) Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries*, **73**, 285–334.

Pinello,L. *et al.* (2011) A motif-independent metric for DNA sequence specificity. *BMC Bioinformatics*, **12**, 408.

Pinello,L. *et al.* (2014) Applications of alignment-free methods in epigenomics. *Brief. Bioinf.*, **15**, 419–430.

Radman-Livaja,M. and Rando,O. (2009) Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.*, **339**, 258–266.

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

Schneider,T.D. and Stephens,R. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Segal,E. and Widom,J. (2009a). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.

Segal,E. and Widom,J. (2009b). What controls nucleosome positions? *Trends Genet.*, **746**, 1–9.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

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

Thompson,J.D. *et al.* (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res.*, **22**, 4673–4680.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267–288.

Tillo,D. and Hughes,T. (2009) G + C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, **10**, 442.

Triola,M. (2012) *Elementary Statistics 12th edn*. Pearson, San Francisco.

Trybula,S. (1947) Some problems of simultaneous minimax estimation. *Ann. Math. Statist.*, **29**, 245–253.

Valouev,A. *et al.* (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.*, **18**, 1051–1063.

Whitaker,J.W. *et al.* (2014) Predicting the human epigenome from DNA motifs. *Nat. Method*, **15**, 390–406.

Yuan,G. (2012) Linking genome to epigenome. Wiley Interdisc. *Rev. Syst. Biol. Med.,* **4**, 297–309.