Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes

Raffaele Giancarlo [a], Simona E. Rombo [a], Filippo Utro [b,*]

[a] Dipartimento di Matematica ed Informatica, Università degli Studi di Palermo, Via Archirafi 34, 90123, Palermo, Italy
[b] Computational Biology Center, IBM T. J. Watson Research, Yorktown Heights, NY 10598, USA

ARTICLE INFO

Article history:
Received 21 February 2018
Received in revised form 14 June 2018
Accepted 30 June 2018
Available online xxxx

Keywords:
Computational biology
Algorithms and complexity
Formal languages
Combinatorics on words

ABSTRACT

Epigenomics is the study of modifications on the genetic material of a cell that do not depend on changes in the DNA sequence, since those latter involve specific proteins around which DNA wraps. The end result is that Epigenomic changes have a fundamental role in the proper working of each cell in Eukaryotic organisms. A particularly important part of Epigenomics concentrates on the study of chromatin, that is, a fiber composed of a DNA-protein complex and very characterizing of Eukaryotes. Understanding how chromatin is assembled and how it changes is fundamental for Biology. In more than thirty years of research in this area, Mathematics and Theoretical Computer Science have gained a prominent role, in terms of modeling and mining, regarding in particular the so-called 10 nm fiber. Starting from some very basic notions of Biology, we briefly illustrate the recent advances obtained via laboratory experiments on the organization and dynamics of chromatin. Then, we mainly concentrate our attention on the contributions given by Combinatorial and Informational Methodologies, that are at the hearth of Theoretical Computer Science, to the understanding of mechanisms determining the 10 nm fiber. We conclude highlighting several directions of investigation that are perceived as important and where Theoretical Computer Science can provide high impact results.

1. Introduction

The genomic DNA of an organism, by means of the genes contained in it, encodes the information dictating the working of a complex biochemical machine whose aim is to produce proteins. In Eukaryotes, it is packed into the cell nucleus forming an extremely compact, three-dimensional (3D, for short) biochemical complex, known as chromatin. Its basic building block is the nucleosome, a spool-like complex composed of histone proteins at the center, with DNA wrapped around. Such a packaging process can achieve staggering 3D volume reduction factors. As well put in [1], geometrically, for Human DNA the reduction factor is as the one required to successfully store a 40 Km (24 miles) long very thin thread into the volume of a tennis ball. The role of chromatin is not limited to such a compression process, as it was initially thought, since it has also a deep influence on gene expression and regulation (simply put, the production of proteins and the regulation of when such a process should start and end for a given gene). In this respect, its role is so characterizing of Eukaryotes that

* Corresponding author.
E-mail addresses: raffaele.giancarlo@unipa.it (R. Giancarlo), simona.rombo@unipa.it (S.E. Rombo), futro@us.ibm.com (F. Utro).

https://doi.org/10.1016/j.tcs.2018.06.047
0304-3975/

gene regulation has a different logic in Prokaryotes, due to the absence of chromatin [2]. These fundamental differences are essential for Eukaryotic organisms to express genes in the great number of diverse patterns that are at the base of their biological complexity.

In order to proceed gradually towards the relevant issues for Mathematics and Computer Science, it is felt as appropriate to give a road map of the paper. With that objective in mind, some key biological facts regarding chromatin structure and function are highlighted in Section 2. It is quite fortunate that the past few years have seen some revolutionary advances and discoveries regarding chromatin structure, in particular with respect to its hierarchical organization. Although we provide a general picture of the State of the Art, we then focus on the first level of that organization, technically, the 10 nm fiber. Indeed, it is in the study of how that fiber is formed that Information Theory and Theoretical Computer Science have had the most impact. Consequently, Section 3 is dedicated to the presentation of State of the Art regarding how to obtain mathematical models that would account for the 10 nm fiber. The need for this type of studies was clearly stated and initiated over 30 years ago by Kornberg in a seminal paper [3], in which it is asked according to which mechanism (formalized via mathematical models) nucleosomes are positioned along a genome. From now on, this latter process is referred to as nucleosome positioning. Section 4 provides an additional level of detail in terms of computational data mining tools specifically designed for chromatin studies on the 10 nm fiber, which reveal a very rich world of "combinatorial messages" associated to chromatin organization and dynamics. Finally, the last Section offers some conclusions and future directions of investigation, highlighting the fact that chromatin studies offer plenty of challenging opportunities for Information Theory and Computer Science to make a real impact on an important area of Science: for instance, very little is known regarding the "combinatorics and mathematics" that could account for the higher levels of chromatin organization recently discovered.

2. The biology of chromatin: basic key facts

We give here an extremely concise presentation of very complicated processes, distilling-off the most important aspects of them in relation to the aim of the paper.

2.1. Function

2.1.1. Packaging the DNA into the nucleus

As already stated in the Introduction, Eukaryotic DNA can be seen as a very long thread. The identification of the processes by means of which this thread is packaged into the nucleus of a cell is a classic and still very active area of research in molecular biology. Relevant is also the identification of the organization and architecture of chromatin. As a matter of fact, thanks to recent technological advances, it is possible for the first time to observe its structure *in vivo*, e.g., [4], rather than via its artificial reconstruction *in vitro*. The result is a sweeping change in the model describing that architecture [4–6], apparently making obsolete classic textbook presentations of the subject. [7].

The basic folding step. The fundamental "brick" of chromatin is the nucleosome (as already stated, a spool-like protein-DNA complex). The distribution of nucleosomes along a genome forms what is known as the 10 nm fiber, which provides a first level of DNA packaging, in the form of length reduction. Because of its centrality for this contribution, we single it out by providing a detailed rendering of its structure in Fig. 1. Although the existence of the 10 nm fiber is an established fact and its organization has been long known [8], recent studies based on nanoscopy and computer simulations have allowed to observe its organization at a single cell resolution *in vivo* [9], i.e., as it really is in a single cell rather than as the result of observations over a population of cells. The important finding reported in the mentioned paper is that nucleosomes forming the 10 nm fiber are organized into interspersed groups. Due to the reminiscence of those groups with egg clutches, they are referred to with the same term. The accuracy of such a model for the 10 nm fiber has been acknowledged and indirectly confirmed by the findings in [4], which have been obtained with the use of an even more advanced technology than the one used in [9].

DNA packaging via a constrained nucleosomic disorder. The formation of nucleosomes along the genomic DNA is not enough to achieve the level of compaction qualitatively indicated in the Introduction. More is needed. Indeed, the classic textbook model [8] that accounts for the condensation of DNA into chromosomes resorts to the use of a series of folding steps, summarized in Fig. 2. Although the model is supported by *in vitro* evidence, it has been elusive to capture *in vivo* [10,11]. In particular in regard to the 30 nm fiber. Indeed, this latter is characterized by a very regular structure of nucleosome arrays arranged on top of each other. The existence of this type of structure *in vivo* has been seriously challenged [4,9], and with it, the *in vivo* accuracy of the textbook model. However, it is interesting to note that the "egg clutches" in Fig. 1 observed *in vivo* correspond to the necklace of pearls in Fig. 2. The new model of hierarchical chromatin architecture that has emerged, e.g. [5,6], is summarized in Fig. 3. The top level is given by the segregation of chromosomes into territories, a fact known since the 80's and well established [12]. Very briefly, chromatin seems to be a disordered fiber that "bends" on itself in 3D to achieve both compaction and proper cell functioning [4,13]. The mechanisms that actually govern this bending so as to obtain segregated chromosomes (see Fig. 3(1), again) starting from the 10 nm fiber (see Fig. 1 and 3(3) again) have not been fully identified yet and their discovery is the object of intense investigation. The state of the art is in [14,15].

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

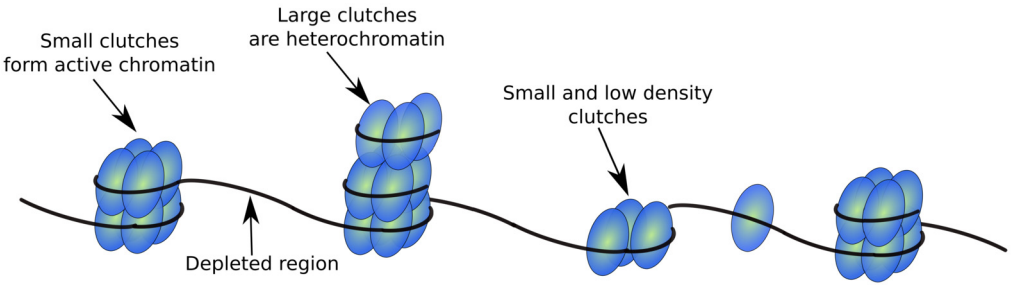

Fig. 1. A representation of how nucleosomes, "the eggs" in the figure, assemble into clutches of different sizes along the genome, the "rod" in the figure. Heterochromatin is the part of chromatin very tightly packed. Active chromatin, also referred to as euchromatin, is the part of chromatin that is lightly packed. The regions of the genome in which there are very few or no nucleosomes at all are referred to as *depleted regions* while the ones where nucleosomes have a high density are referred to as *enriched regions*. Both types of region can be of various lengths.

2.1.2. Regulating gene expression

Following [1], a gene is a portion of DNA that encodes for a protein or an RNA molecule. At a very high level and informally, the process that transforms a gene into a protein consists of two steps. In the first step DNA is transcribed into RNA, then the RNA molecule is translated into a protein. The interested reader can find details of the two processes in [1]. In case the gene produces only an RNA molecule, the process of translation does not take place. Those genes are referred to as *non-coding*. In both cases, the described process is referred to as *gene expression*. Chromatin organization and its core components, i.e., histones, play a key role in this process.

A simplistic point of view, resorting to notions well known in computer science, is as follows. Gene expression can be seen as a transduction process performed by a sort of Turing machine, which has the DNA sequence of the gene on the input tape and which produces the corresponding sequence of the gene product on the output tape, usually corresponding to a sequence of amino acids. The first position of the input tape corresponds to the so called Transcription Start Site (TSS for short), i.e., the DNA letter from which the transduction process must start. In order for this process to work properly, the gene DNA sequence must be accessible, in particular the TSS. However, a look at Fig. 1 immediately reveals that the part of DNA wrapped around the histones, i.e., the DNA that forms nucleosomes, is not accessible. Therefore, if part of a gene is covered by nucleosomes, it cannot be transduced properly. The point is that, in order for a gene to be expressed, nucleosomes occluding it in full or part must be moved elsewhere.

A more accurate rendering of how chromatin affects gene expression is beyond the scope of the introductory material presented here and needed for this study. However, for completeness, it is important to mention that chromatin can be divided in two types: euchromatin and hetherochromatin (see Fig. 1 again). The first type is composed of lightly packed DNA and it is usually associated with genomic regions under active transcription. The second type is composed of tightly packed DNA. Usually it is associated with transcriptionally inactive genomic regions. Its fundamental role in the proper working of a cell is well presented in [16]. Given the body of work that the area of string algorithms has produced regarding repetitive structures in strings, also with attention to genomics, e.g., [17,18], we mention that highly repetitive genomic regions are a threat to cell stability. Hetherocromatin has a fundamental role in suppressing the events that "unguarded" repetitive areas of the genome would trigger.

Finally, it is also important to mention that histones, a building block of chromatin, play a fundamental role in gene regulation via families of modifications. Intuitively, those latter can be seen as proteins that are obtained by histones via modifications that occur once an histone protein has been obtained via the process of translation. Such changes are referred to as *Post-translational Modifications*. The interested reader can find an excellent presentation of this topic in [19].

2.2. The 10 nm fiber: mechanisms for nucleosome positioning and dynamics

Even from the brief description given in the previous Section, it is evident that chromatin is far from being a static object, although its organization can be described via a static hierarchical model. In particular, nucleosomes must form along the genome and they must "move around" to make accessible the parts of the genome that are needed to be accessible at any given time for the proper functioning of the cell. Mentioning once again that the mechanisms of chromatin dynamics at any scale are still in the process of being discovered, (see again [14,15] and [20]), we focus on the 10 nm fiber, in particular on nucleosome positioning. It is worth to state the problem considered here by citing verbatim Radman-Livaja and Rando [21]: "The remarkably uniform and conserved nucleosomal organization of gene promoters begs the question: what determines nucleosome positions throughout the genome? Are nucleosome positions primarily "encoded" in the DNA sequence (cis factors) or are they a consequence of the regulatory activity of chromatin remodelers, transcription factors and the transcription machinery (trans factors)?" Remodelers and transcription factors (TFs, for short) are proteins that bind to specific DNA positions and, in doing so, intuitively and with respect to the DNA position where the binding takes place, they either "push" nucleosomes elsewhere or "get in the way" of their formation there. We next highlight some fundamental findings regarding the two aspects characterizing the stated problem. For completeness, we mention that they

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

The Classic Model of DNA Packaging Into a Chromosome.

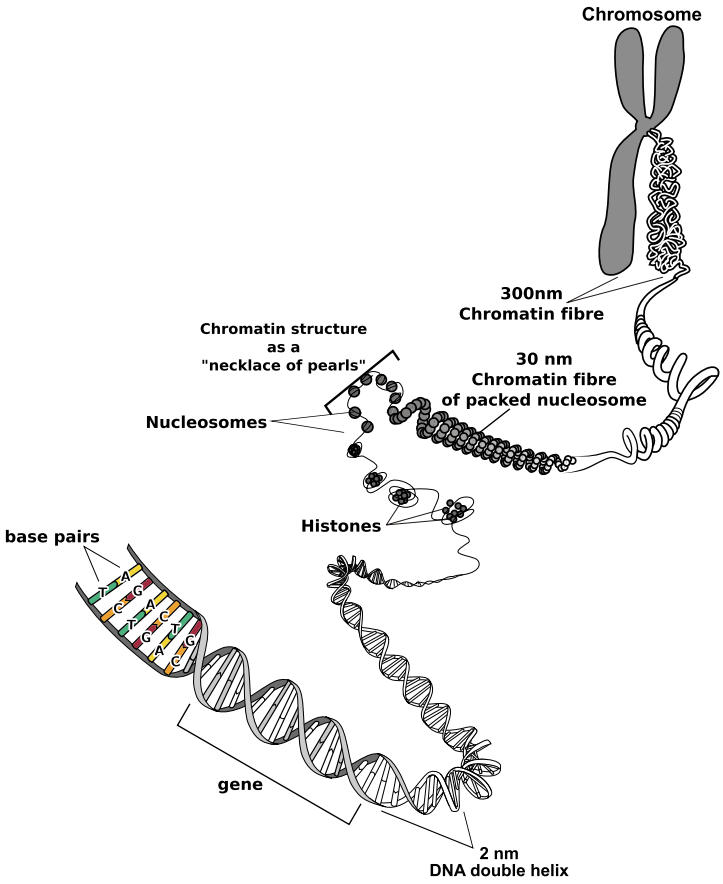

Fig. 2. The classic model accounting for DNA packaging process at the chromosomal level.

The key steps of this process are shown in Fig. 2. It is worthy of mention that only the first step of this process is known to take place *in vivo*, while for some of the others there is only evidence *in vitro* or they are simply hypothesized. The following explains in more detail the representations shown in the figure. a) The DNA double helix. b) The necklace of pearls, which is obtained through the wrapping of DNA around a series of spool-like structures, each of those latter obtained via a proper assembly of histones. Each spool, with DNA wrapped around, is referred to as a nucleosome, which is "the pearl" of the necklace. c) the 30 nm fiber is observed *in vitro* and it is obtained by the packaging of nucleosomes into regularly spaced arrays. d) The packaging goes on to form chromosomes.

are not mutually exclusive: as a matter of fact, they may very well coexist [22], although the extent of which each factor contributes to nucleosome positioning is still unknown.

2.2.1. Intrinsic positioning instructions "encoded" into the genome

As all the proteins that bind to DNA, histones are "choosy" with respect to the DNA sequences "they like to bind to". The existence of those biochemical histonic "like" and "dislike" votes has been known for a long time, e.g., [3], and they have been the object of intense research [21]. However, such a fact alone is not enough to convincingly claim that the genome has "encoded" in its sequence a map of where nucleosomes must be or not. The design of an experiment giving evidence of such a fact is challenging. Quite remarkably, with the advent of the new DNA sequencing technologies and following ground-breaking work in [23], Kaplan et al. [24] have succeeded. Their experiment is worth a brief conceptual description.

It is known how to construct a nucleosome occupancy map *in vivo*. That is, a numeric vector in which each index corresponds to a genomic position and the entry at that index gives a confidence score, referred to as occupancy value, that such a position is covered by a nucleosome. An example is provided in Fig. 4, e.g., top curve. However, since in order to form nucleosomes, histones have to bind to DNA, they have to face the "competition" of TFs and remodelers that,

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

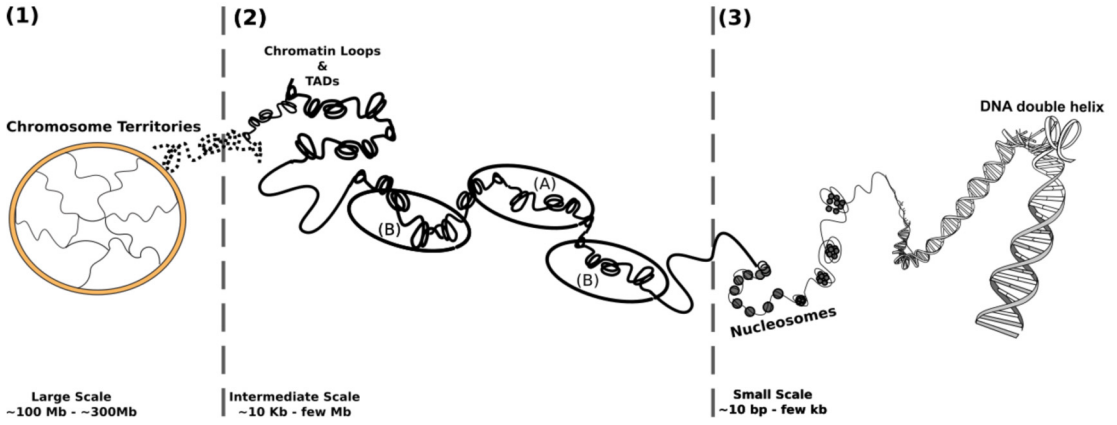

Fig. 3. A novel model summarizing chromatin organization into the nucleus hierarchically.

(1) The nucleus divided into chromosome territories (only boundaries shown). That is, each chromosome, packed into chromatin, is confined in a particular part of the nucleus and, in 3D, there is very little contact (over time) with other chromosomes. The scale refers to the length of the genome confined into each territory. (2) Each chromosome can be further divided into two parts: Active compartments (denoted by (A) in the Figure) and inactive compartments (denoted by (B) in the Figure). There is, over time, very little contact occurs between chromatin (the rod) in compartments (A) and (B). At the same scale as compartments, there are "local parts" of the genome that have contact in 3D., i.e., the genome forms loops that "cluster" together, with very little interaction among different "clusters". Those latter, technically, assume the name of Topologically Associated Domains (TADs). (4) Finally, we have the 10 nm fiber, whose architecture is provided in Fig. 1.

on occasions, may cause nucleosome formation in places where histones do not have high preference for the underlying sequence. Therefore, the *in vivo* map would also account for the contributions of remodelers and TFs to the positioning of nucleosomes, since those latter would be present in the culture of cells from which the map is derived. Kaplan et al. have constructed an *in vitro* nucleosome occupancy map with the property of being of high quality, and in which experimental conditions have favored that histones binding position do not depend on the competition of remodelers and TFs, since those latter are not part of the purified DNA from which the map is obtained. Quite remarkably, there is a very high correlation between the *in vivo* and *in vitro* maps for *S. cerevisiae*, indicating that indeed the underlying DNA sequence has an influence in the determination of nucleosome positions on a genomic scale. An example of such a correlation is provided in Fig. 4 (top two curves). So, very informally, the genome "encodes" the positioning of nucleosomes along itself. We anticipate that the mathematical nature of such an "encoding" is discussed in Section 3.2.

2.2.2. Moving nucleosomes around along the genome

As already stated, all of DNA is packaged into chromatin and, for each cell, chromatin organization follows the model described in Fig. 3. However, it has been very well known for a long time [25] that such a layout is not maintained exactly throughout different cell types in the same organism and in all parts of the genome within each cell, although the underlying DNA sequence is the same. That is, different cells may have nucleosomes in differing genomic positions.

In order to give an account of those changes, the main mechanism to consider is the action of histone competitors: remodelers and TFs. Indeed, remodelers can dislocate nucleosomes from their positions and TFs compete for the same binding sequences. A legitimate question to ask is whether such a competition is well documented or speculative. In this respect, there is quite some work in progress, however most of the results indicate, at the qualitative level, that it has an impact on chromatin dynamic changes, e.g., [26,27]. For our purposes, it suffices to highlight that the *in vitro* map mentioned earlier is not always in good agreement with the *in vivo* map. In particular, the absence of nucleosomes in genomic coding regions measured *in vivo* increases with the expression level of the associated genes with respect to nucleosomal presence measured *in vitro*. These results indicate that TFs, chromatin regulators and active transcription influence the resulting nucleosome organization *in vivo* [28].

As an important exemplification of the dynamic process and competition just outlined above, we concentrate on one particular aspect of it: the creation of Nucleosome Free Regions (NFR, for short) around TSS. Indeed, the understanding of the mechanisms that contribute to create a NFR would highlight one of the most basic steps of chromatin dynamics and, consequently, of gene expression and regulation.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

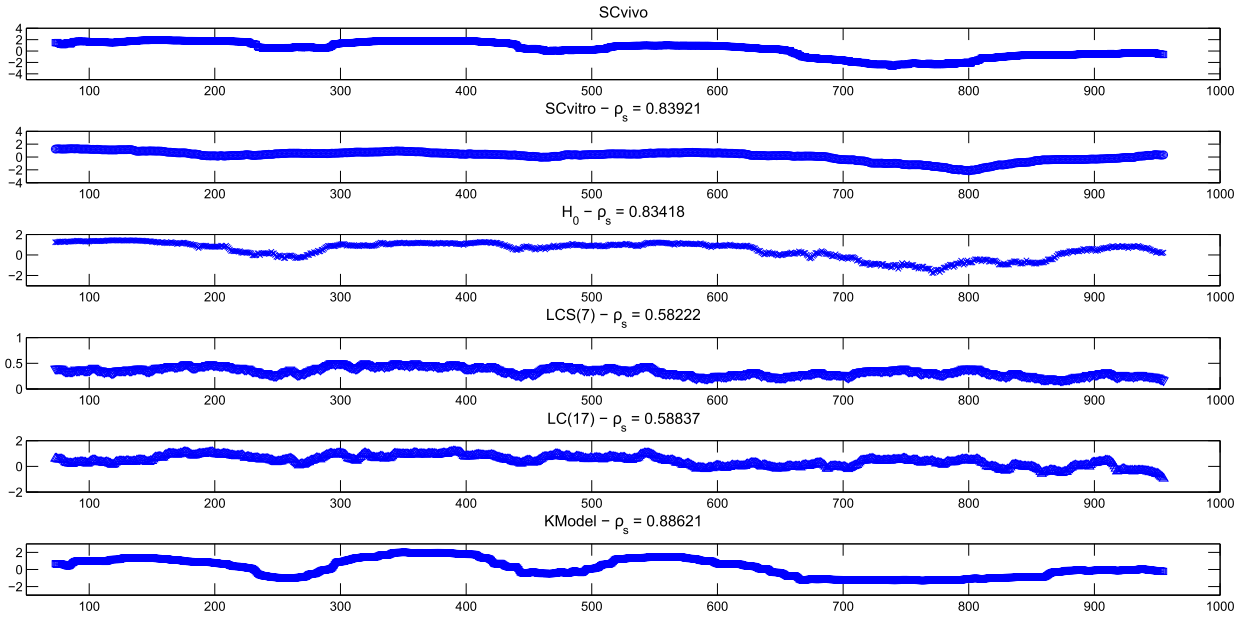

Fig. 4. (Top two curves) For a genomic region of 1024 base-pairs in length, the plots of the occupancy values for *S. Cerevisiae in vivo* and *in vitro* (`SCvivo` and `SCvitro`, for short), respectively. The abscissa indicates a position in the sequence and the corresponding ordinate provides the occupancy value. (The remaining plots) They provide curves analogous to the ones just described. However, the "occupancy" value of a position has been determined by one of the computational methods considered in Section 3.2.2 and according to the sliding window method described in that Section, except for `Kmodel` which has been used without a sliding window. Taking as reference `SCvivo`, the value of the Spearman rank correlation coefficient is also indicated for each curve.

As already mentioned, once that DNA is packaged into chromatin, genes may not be accessible, in full or in part (see Fig. 1 again). In order to obtain the product(s) of genes that are not accessible, the chromatin organization around them has to be changed. Such a task is accomplished by chromatin remodelers. Even before the transduction process involving a gene begins, the TSS upstream of it is made accessible via the creation of a NFR around it. The structure of that region is depicted in Fig. 5(a). For quite some time, the creation of such an NFR has been attributed to the stiffness of the so-called poly(dA:dT) tracts, i.e., the "combinatorial message" A^k or T^k, for some constant k. The rationale being that their poor bendability favors the creation of unstable nucleosomes that are then easy to dislocate. The main experimental evidence supporting this thesis were the structural characteristics of the poly(dA:dT) tracts and the abundance of A-T nucleotides in NFRs [29]. Therefore, under this thesis, poly(dA:dT) plays only a passive role in determining nucleosome positioning: via its stiffness, its relative abundance levels along the genome would demarcate genomic regions fit for nucleosome occupancy or not. Recently, such a point of view has been challenged and, to some extent, changed. Indeed, Lorch et al. [30], among others, have provided experimental evidence supporting the thesis that NFRs are created by an active mechanism of nucleosome rearrangement due to chromatin remodelers rather than by a passive mechanism connected to sequence composition. Quite remarkably, key elements of such an active mechanism have recently been described in detail via *in vitro* experiments by Krietenstein et al. [20]. Those findings are reported in Fig. 6. Even more remarkably, the poly(dA:dT) tract assumes the real role of a message. Indeed, it must be recognized by the chromatin remodeler RSC in order to start the removal of nucleosomes to form an NFR. Therefore, the DNA sequence actually guides the chromatin remodelers in creating the NFR. However, so far, in this very important setting, only the message poly(dA:dT) has been identified.

For the interested reader, it is worth of mention that the latest account of the many aspects of chromatin remodeling is reported in [31], where remodelers are described as nano-motors acting on nucleosomes. A mechanism by means of which those motors are placed in the right place and activated at the right time has also been hypothesized. It involves interactions between histones and remodelers via some kind of "message passing".

3. The mathematical and informational nature of chromatin organization: simple, genome-wide, closed-form formulas for the 10 nm fiber

Before moving on, it is useful to recall three main facts from the previous section.

(a) DNA is packaged both *in vivo* and *in vitro* into chromatin. In terms of nucleosomes, bulk chromatin is composed of arrays of regularly spaced nucleosomes *in vitro*, while such a regularity is somewhat lost *in vivo*.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

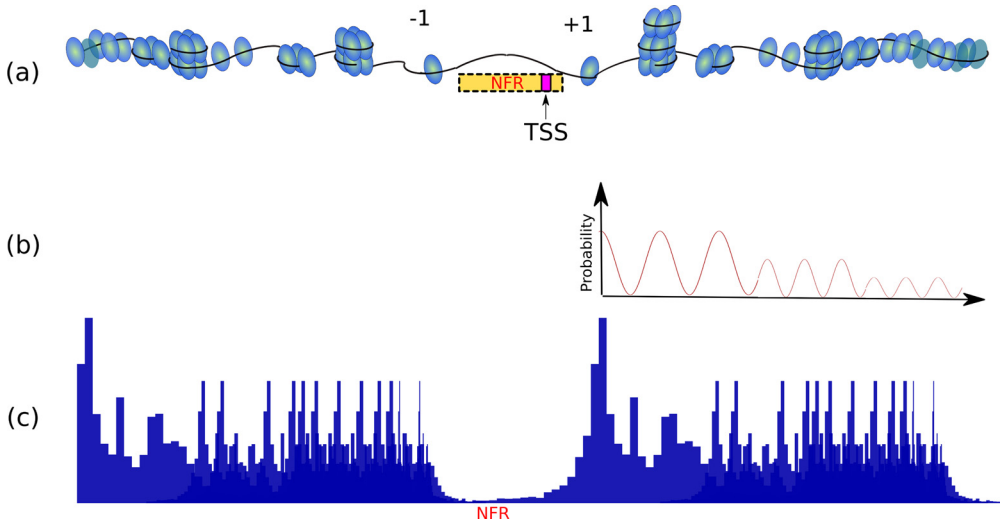

Fig. 5. (a) The ideal representation of a NFR upstream of a gene that is "ready" for transcription. The TSS is accessible to the "machine" responsible for carrying out the transcription. The NFR is delimited upstream and downstream by nucleosomes, with positive numbering towards the so-called 3' end of the gene and negative in the other direction. (b) The curve of the probability distribution of nucleosome occupancy predicted by the so-called "barrier model". The barrier is made by the nucleosome clutches numbered $+1$ in the 3' direction and -1 in the opposite direction. (c) a histogram of nucleosome occupancy coming from sequencing experiments in *S. Cerevisiae* that confirm, on a genomic scale, the validity of the "barrier model". In the histogram, the peacks correspond to positions that have been found to be occupied by nucleosomes in 4799 TSS in *S. Cerevisiae*. The histogram has been obtained by aligning and then superimposing those 4799 occupancy maps [32].

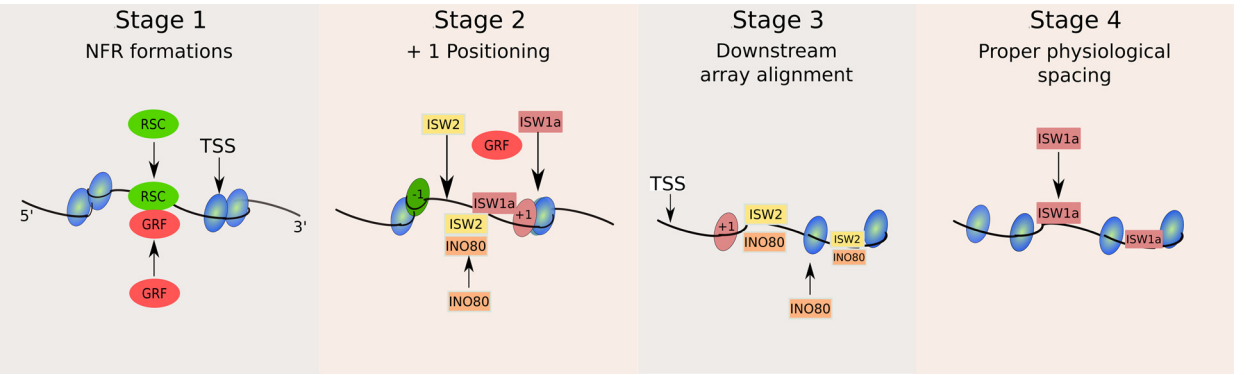

Fig. 6. A representation of the four stages of NFR formation and nucleosome (re)positioning in Yeast. In stage 1, a remodeler (RSC) generates a NFR by removing nucleosomes. The action of RSC starts either because it is recruited to the task by another remodeler (GRF) or when it "reads" a properly positioned poly(dA:dT) "message". In stage 2 ($+1$ positioning), the $+1$ nucleosome is positioned by the action of three remodelers (INO80 ISW2, ISW1a). In stage 3 (array alignment), ISW2 and/or INO80 generate nucleosomal arrays downstream the $+1$ nucleosome with a "long" nucleosomal spacing, i.e., the distance in base-pairs between two consecutive nucleosomes. In stage 4 (proper spacing), ISW1a generates proper nucleosomal spacing in downstream nucleosome arrays.

(b) From the biological point of view, the genome sequence alone has enough information to determine nucleosome occupancy along the genome. The term information in the previous context means that nucleosome occupancy is influenced by the underlying genome.
(c) Nucleosome positions change in response to the needs of a cell. In particular, in order to allow for the expression of genes, NFRs are created via the cooperation of several remodelers.

It is very natural, and very important, to establish the mathematical nature of those facts, i.e., whether or not there exist mathematical models that can account for them, even in part. Quite remarkably, the answer is positive for (a) and (b) via some very simple formulas. It is wide open for (c), together with many other mathematical modeling problems relating to the 3D chromatin organization depicted in Fig. 3, which will be mentioned in Section 5.

From now on, we concentrate on Information-theoretic and algorithms on words that are relevant for the topics outlined above. However, for the interested reader, a full account of computational resources for nucleosome studies is provided in [33]. The specific problem of histone modifications identification on a genomic scale, and as it is solved with the use of Machine Learning techniques, can be found in [34] and references therein (see also [35] for a review of the subject). Moreover, computational methods that provide a genome-wide landscape of protein-DNA interaction, in particular modeling

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

TFs and histone binding competition, have been recently received attention in the Literature, e.g., [36]. Once again, they are based on machine Learning and techniques of Data Integration from several sources.

3.1. The barrier model: a regular packaging via a random process

Kornberg and Stryer [25] have shown that the regular spacing of nucleosomes in bulk chromatin can be explained via a stochastic process. The intuition is as follows. Nucleosomes must obey steric constraints, i.e., only a single nucleosome can occupy a certain amount of "space". Or, put it very informally, nucleosomes cannot be on top of each other. Assume now that a nucleosome is positioned before the others in close proximity of a barrier, e.g., a physical location where nucleosomes cannot be. Typical examples of nucleosome barriers may be sequences disfavouring nucleosome positioning and DNA binding proteins (e.g., TFs) bound to specific locations in the genome (again, steric constraints to satisfy), or even "road block" nucleosomes: nucleosomes that are particularly well positioned and difficult to evict [37]. For the sake of discussion, let us assume that the nucleosome is to the immediate right of the barrier. Then, its positioning there influences the chances of the positions following it being occupied by a nucleosome. Intuitively, those chances are influenced to a lesser and lesser extend as one moves away from the barrier.

Formally, a DNA sequence is seen as a one dimensional line, with a beginning point (the barrier). Moreover, each segment of that line can be labelled in two ways: bead (if covered by a nucleosome) or bare (in the other case). Now, the following formula gives the probability of a position t units from the barrier to be bare:

$$p(t) = s(t-1)w(1,0). \tag{1}$$

Qualitatively, s gives the probability mass obtained by considering the probabilistic contribution given by each possible arrangement of beads in the positions preceding t, while $w(g,n)$ gives the probability of having n beads on a segment of length g. In the case of interest, the number of beads is zero and the segment length is one. Quantitatively, both s and w can be expressed in terms of closed formulas, which are very easy to compute numerically. The interested reader can refer to [25] for details. Obviously, the probability of a position t units from the barrier to be bead, i.e., occupied by a nucleosome, is $1 - p(t)$. An example of the graph of this latter curve is depicted in Fig. 5(b). In that Figure, a barrier is to the immediate left of the $+1$ nucleosomes delimiting the NFR.

Kornberg and Stryer, in their original study, validated the proposed stochastic model by showing that it is in agreement with experimental data. However, a genome-wide validation of it has been obtained only recently by Mavrick et al. [32] for yeast. The experimental data was obtained by extracting, from a yeast nucleosome occupancy map produced by the authors, the occupancy data limited to the genomic regions around (4000 bp window) each of 4799 known TSS genomic locations. The data so extracted were then aligned on the TSS and binned to obtain an histogram, shown in Fig. 5(c), with the intent of summarizing a common occupancy pattern, if any. Quite remarkably, the histogram shows a pattern as the one predicted by the Barrier Model. In particular, according to that study, the $+1$ nucleosome acts as if it were the barrier in the Kornberg and Stryer model, with respect to the nucleosome distribution downstream of it. The same assertion holds for the -1 nucleosome regarding nucleosomes upstream of it. In a sense, those two nucleosomes act as the above mentioned "road block" nucleosomes. However, in view of how the NFR is created (see again [20]), the concurrent presence of remodelers in their proximity is not to be excluded, although the data in [32] are not amenable for such a more in-depth investigation.

It has to be mentioned that such a probabilistic model has been generalized and validated by Möbius and Gerland [37]. In that study, it has also been investigated to what extent possible sequence preferences affect the pattern observed in [32]. Moreover, a quantitative analysis is provided of yeast nucleosome positioning data, both up- and down-stream from NFRs. In this respect, the result is that, although the typical patterns on the two sides of the NFR are different, they are both quantitatively described by the same physical model with the same parameters, but with different boundary conditions.

Other experimental studies provide evidence that in more complex eucaryotes, such as human, many fewer nucleosomes are positioned according to the Barrier Model than for example in yeast [38]. This is possibly due to the massive size of genes resulting in much more space between the potential barriers located at promoters, enhancers, insulators, and regulatory regions.

3.2. The combinatorial and information theoretic nature of the nucleosome occupancy "encoded" in the genome sequence

In contrast to the Barrier Model accounting for the packaging of DNA, the mathematical nature of the sequence specificity mentioned in (b) has been a bit more challenging to be (at least in part) formalized and experimentally validated on a genomic scale. Indeed, if one is interested in machine learning procedures that, properly trained, can predict reliably nucleosome occupancy at a genomic scale, then there are plenty. The interested reader is referred to [39] for the State of the Art. However, if one is interested in closed-form mathematical formulas or even universal mathematical laws of modeling sequence specific nucleosome occupancy, there has been even some skepticism in regard to their existence, e.g., [40]. Somewhat surprisingly due to the simplicity of the approach, those formulas do exist, as shown in [41]. They are based on very well known and very basic measures of the complexity of a sequence. We present next the work in [41].

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

3.2.1. Sequence complexity measures

The complexity of a sequence can be formally defined by resorting to techniques and ideas coming from the following related areas: sequence combinatorics and linguistic complexity [42,43], Shannon information theory [44], and Kolmogorov–Chaitin algorithmic complexity [45]. In terms of actual numeric evaluations, measures stemming from this latter area are not computable and can only be heuristically approximated via data compression programs (see [46,47]). However, the corresponding heuristics offer no performance guarantee on how good the corresponding approximations are. Therefore, only measures obtained from the first two mentioned areas are used in [41], as follows.

Combinatorial and Linguistic. Let Σ be a finite alphabet of symbols. Given a sequence x of length n, defined over the alphabet Σ, and an integer $i \leq n$, let $LCS(i)$ be the number of distinct subsequences of length i that are present in x, normalized by n. Now, fix an integer $k \leq n$ and let:

$$LC(k) = \sum_{i=1}^{k} LCS(i).$$

A few remarks are in order. LC is a normalization of a measure due to De Luca and Varricchio [43] and it is related to the linguistic sequence complexity introduced by Trofonov [42]. Both LC and LCS measure the complexity of a sequence based on how many distinct subsequences are present in it. It is well known that the lower that number, the less complex the sequence is.

Information Theoretic. The empirical entropy H_0 of a sequence x is defined as follows:

$$H_0(x) = -\sum_{i=1}^{|\Sigma|} \frac{n_i}{n} log_2 \frac{n_i}{n},$$

where n_i is the number of occurrences of symbol a_i in x. It is worthy of mention that there is an important difference between empirical entropy and the entropy defined in a probabilistic setting [44]. Indeed, as detailed in [48], Shannon entropy is an expected value taken on a probabilistic process that may emit a possibly infinite ensemble of sequences, while empirical entropy is defined point-wise for any sequence: it measures the amount of information needed to optimally encode it, without any reference to "a probabilistic model" generating sequences. That is, it is a punctual and intrinsic measure of information characterizing a sequence alone, rather than a measure of uncertainty characterizing a model generating sequences. Also in this case, it is well known that the lower the value of H_0, the less complex the sequence is.

3.2.2. Experimental validation

We need to briefly present how to obtain maps *in silico*, in terms of a generic algorithm $\mathcal{A}$ that takes as input a sequence and produces a non-negative number as output. Assume that the genome of which the map has to be built is divided into maximal regions of contiguous bases $R = \{[s_1, e_1], [s_2, e_2], \ldots, [s_q, e_q]\}$, where the interval endpoints naturally indicate the start and end genome coordinate of each region, respectively. Each region in R is swept, from left to right, by a window of length 147. The corresponding sequence is given in input to $\mathcal{A}$ and the value returned in output is assigned to the genomic position aligned with the center of the window, e.g., when $[s_1, e_1]$ is swept, the result is a map for the genomic positions in $[s_1 + 73, e_1 - 73]$. The window size has been chosen to coincide with the length of a DNA sequence that "wraps around" histones to form a nucleosome.

In order to show that simple complexity formulas can predict nucleosome occupancy with a high degree of reliability, an extension of the procedures followed in [24] has been proposed in [41]. That is, for a given organism, a genome-wide complexity map is created (via the in silico procedure described above) and for each of the measures described in Section 3.2.1. Then, the correlation between each complexity map and a nucleosomal map for the given organism is computed. To this end, four maps of model organisms have been used in [41]. The first three are common to the study by Kaplan et al. [24]. Namely, the normalized *in vitro* and *in vivo S. cerevisiae* maps, the adjusted occupancy *C. elegans* map by Valouver et al. [40] (chromosome 2), and the *D. melanogaster* maps by Mavrich et al. [49] (chromosome 2).

Fig. 4 shows the *in silico* maps obtained with the use of algorithms that compute empirical entropy and linguistic measures of part of the genomic sequence of *S. Cerevisiae*. For those latter measures, the chosen values of k are the best performing in terms of correlation with nucleosome occupancy maps of the organisms included in the study in [41]. For comparison, a leading machine learning algorithm is also reported in the Figure, i.e., `Kmodel` [24]. The mentioned Figure also gives an example of the excellent correlation among the *in vivo* and *in vitro* and the *in silico* maps. More in general, a good level of correlation, genome-wide, has been shown [41] between each of the maps of the selected organism and each of the corresponding complexity maps. The fact that `Kmodel` produces maps in good agreement with the ones of *S. Cerevisiae* and *C. Elegans* can be inferred from results in [24].

The results just outlined have important implications. First, they give a mathematically rigorous form to the "information" about nucleosome positioning "encoded" in a genome. Second, from the methodological point of view, they close an important gap between the Barrier Model accounting for DNA packaging and sequence-specific occupancy, which was lacking

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

analogous elegant closed-form formulas. In algorithmic terms, the complexity measures can be efficiently computed using standard C++ data structure and core functions provided by the SeqAn library [50] to build and traverse a suffix trie [51]. Quite remarkably, the procedures so obtained are at least an order of magnitude faster that `Kmodel` and need no training.

Finally, it is well known that entropy estimation, in particular for DNA sequences, is a very rich area of investigation. The interested reader may find relevant methodologies in [46,47,52]. Therefore, it is quite natural to ask whether known techniques would bring better results with respect to the very simple empirical entropy used here, in particular when one resorts to compressive estimates of entropy. That is, estimation via the use of data compression programs. This possibility has been considered in [41]. In particular, XM [53] and Arithmetic Coding [54] have been used. The first data compressor is among the best for DNA sequences. The second is quite effective and also offers the advantage to have specific parameters that control how fast the compressor learns statistics about the sequence to be compressed. Although the results of the experiments were good, they were no better than the ones obtained with the use of H_0. It has also to be remarked that the use of higher order empirical entropy gave much worse results than the memoryless case.

4. Genome-wide mining of nucleosomal maps for the identification of specific positioning motifs: combinatorics and algorithmics on strings at work

Algorithms and computational techniques for the identification of motifs, i.e., short regular expressions, in biological sequences has become one of the pillars of data mining in biology and there has been revived interest in it with the advent of the new sequencing technologies. The interested reader can find a good introduction to this, by now classic, topic in [55] and a more recent account of the State of the Art in [56].

From that vast area, a particular branch of interest here concerns the use of statistical scores for the identification of "unusual" sequences within a set of biological sequences. There are two main approaches that have been proposed. One in which a set of patterns and texts are given, possibly with a background probabilistic model that characterizes the "source" emitting the text sequences, and the procedure returns a score for each pattern indicating "how unusual" (with respect to the background model) is the occurrence of that pattern within the texts. An introductory description of the basic techniques, as well as software, is available in [57]. The other, in which, given a sequence, all of its subsequences are assigned a score assessing "how unusual" their occurrences are. A notable example of those methods is given in [58].

To the best of our knowledge, none of those techniques has been applied to the identification of specific motifs and regularities that would characterize nucleosome enrichment and depletion. Indeed, machine learning techniques trained on nucleosomal maps have been used to identify key sequence features. Specifically, the sequence motifs and regularities that have been identified are: (a) the 10 bp periodicity of the dinucleotides AA/TT/TA that oscillate in phase with each other and out of phase with a similar periodicity of the GC denuclotides [23]; (b) poly(dA:dT) tracts [29] (c) the G+C content of a genomic region, with its A+T content also playing some role [59,60]. In terms of specific k-mers, studies about their role in favoring nucleosome positioning have only identified a handful of them [59,60].

Although it remains open to establish to what extent the statistical-combinatorial techniques mentioned earlier can be applied in this context, we now outline an approach reported in [61] in order to identify k-mers that are characteristic of nucleosome depletion and enrichment. However, it is worth mentioning that the technique is general as it can be applied to other tasks in which one has to extract relevant k-mers characterizing one of two biological states or functions.

4.1. Epigenomic dictionary: basic definitions

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 $s = +/-$ means that x is a "characteristic/significant feature" of favouring/disfavouring nucleosome positioning regions in a genome. 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 $\mathcal{D}_{k_{\max},\alpha} = \bigcup_{k=1}^{k_{\max}} \mathcal{D}_{k,\alpha}$ 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.

For a given organism, the weighted dictionary that can be built directly from genome-wide nucleosome positioning maps 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. It is worth mentioning that the efficient collection of k-mer statistics over sets of sequences is essential in order to construct weighted dictionaries on a genomic scale and in a reasonable amount of time. The interested reader can find some recent advances on the collection of those statistics in [62].

4.2. The choice of a weighting scheme for nucleosome positioning

Let E and D be two sets of sequences that are nucleosome enriched and depleted, respectively. 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 E and D, i.e., such a difference in frequency is not due to chance.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

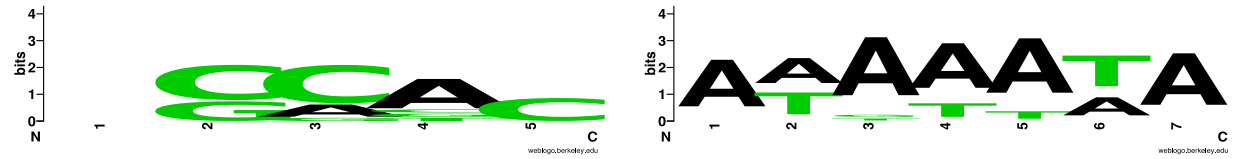

Fig. 7. Sequence Motifs obtained as described in the main text. The one on the left is associated to nucleosome enrichment, while the one of the right to nucleosome depletion.

As elaborated next, (a) can be formalized via Binary Classification in Machine Learning, while (b) via Hypothesis Test in Statistics. The formalization of (a) and (b) in the context of weighted k-mer dictionaries is due to [61], although the Binary Classification technique has been implicitly used by [59] in extracting sequence nucleosome positioning signals in *S. cerevisiae*. For completeness, it is worth mentioning that the identification of an appropriate weighting scheme is strongly related to *feature selection* in Machine Learning [63], although neither of the two schemes outlined next can be regarded as a feature selection technique. This remark poses the problem of investigating those latter in the context of weighted dictionaries.

4.2.1. 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 [64]. 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 (Area Under the Curve) 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 in order to be included in the dictionary.

4.2.2. A weighting scheme based on hypothesis test

Let $\mathbf{Q}$ and $\mathbf{P}$ 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., [65] for the definition of z-score and its uses in data normalization). In order 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 [66–68]. 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, e.g., $1,000$, of times, 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.

4.3. Extracting nucleosome favouring/disfavouring motifs from a collection of dictionary

Once that a dictionary has been obtained, one can build sequence motifs, favouring or disfavouring nucleosome formation. For a given threshold $\beta \geq \alpha$ use it to extract nucleosome favouring motifs, the following procedure has been proposed [61].

(a) For $k \in [1, k_{\max}]$, extract each k-mer from $\mathcal{D}_{k_{\max},\alpha}$ that has a $+$ sign and a threshold at least β to obtain a set of sequences ED. It is worth pointing out that ED may be composed of k-mers of different lengths.
(b) Partition ED into clusters with the use of a clustering algorithm suitable for sequences, e.g., DNACLUST [69], with a sequence similarity threshold of 75%, computed via standard semi-global alignment.
(c) For each cluster, align the k-mers in it via a multiple sequence alignment program, e.g., CLUSTALW [70]. For each alignment so obtained, use a tool that extracts motifs from multiple alignments, e.g., WebLogo [71].

Fig. 7, taken from [61], provides an example. The motifs have been obtained via the construction of a dictionary, Binary Classification weighting scheme, for each of the following model organisms: fly, human, yeast and worm. Then, the intersection of those dictionaries has been taken and the procedure outline earlier for the extraction of favouring/disfavouring nucleosome formation motifs has been applied to the resulting set.

4.4. Outline of experimental findings

Dictionaries have been constructed with the use of both of the mentioned weighting schemes outlined in Section 4.2 for fly, human, yeast and worm [61]. A summary of the main findings reported in that paper, in terms of data mining methodologies, and as well as specific contributions to Epigenomics, are outlined next.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

Table 1
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

Both weighting schemes and the notion of dictionary defined in the preceding Sections provide two very different, and equally powerful, data mining tools. Binary Classification gives a high level description of the most relevant differentiating motifs between nucleosome enrichment and depletion while Hypothesis Test provides a rich level of detail. Therefore, the use of both gives a hierarchical view of the essential sequence features involved in enrichment and depletion.

As for the contributions to Epigenomics, it is worth to recall that "a genomic code" for nucleosome positioning has been claimed to exist [72], although its exact identification has been very elusive so far. The use of the Hypothesis Test weighting scheme gives, at least qualitatively, an insight on how complex such "set of instructions" is. Indeed, Table 1 reports some statistics about the clusters extracted from the dictionaries for the already mentioned organisms, indicating that many different favouring/disfavouring sequence signals are involved in chromatin organization. The interested reader can find additional detail about the role of specific sequence motifs in [61].

5. Conclusions and future directions of research

Here we provided a roadmap of the State of the Art of Mathematical and Computer Science aspect of the chromatin organization characterization. Starting from some very basic notions of Biology, we have highlighted the role that in chromatin studies has been given by two of the main areas of Theoretical Computer Science: Combinatorial and Informational Methodologies. Moreover, the connection of these two areas to Machine Learning is also receiving attention and hopefully it will result in further unifying principles and methodologies, with impact in Life Sciences. As it is evident from the presentation given, this entire subject area offers many interesting research directions, which we outline next.

(1) **Sequence Complexity and Linguistic Composition of Sequences.** Sections 3.2 and 4 have provided combinatorial and algorithmic methodologies that characterize nucleosome positioning. As of now, they lack the specificity regarding genomic hallmarks such as TSS, regulatory and coding regions. It would be of great interest to characterize those hallmarks by extending the mentioned techniques.
(2) **Characterization of Histone Modifications.** This is an important topic, with many implications for Medicine, in particular cancer research. As mentioned, the main known techniques to identify those modification genome-wide are based on Machine Learning paradigms. It would be extremely interesting for Theoretical Computer Science to show that there are algorithms based on combinatorics on words able to perform the same task. It would be even more relevant if, via those algorithms, one would get compositional insights into this important biological process.
(3) **From the Genomic Sequence to 3D Chromatin Organization.** The methods in Section 3 present the mathematical formulas that can "predict" nucleosome architecture in bulk chromatin or at the level of the 10 nm fiber. However, no satisfactory mathematical model is present that reliably accounts for the 3D chromatin organization based only on the genomic sequence. Some work in that direction has been recently done in [73], but this area needs much more attention.

Funding

Istituto Nazionale di Alta Matematica "F. Severi" - GNCS Project 2017 "Efficient algorithms and techniques for the organization, management and analysis of Big Data in the biological context" (to R. Giancarlo and S.E. Rombo).

Istituto Nazionale di Alta Matematica "F. Severi" - GNCS Project 2018 "Processing and analysis of Big Data modeled as graphs in different application contexts" (to R. Giancarlo and S.E. Rombo).

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson, Molecular Biology of the Cell, 4th edition, Garland, 2002.
[2] K. Struhl, Fundamentally different logic of gene regulation in eukaryotes and prokaryotes, Cell 98 (1999) 1–4.
[3] R.D. Kornberg, The locations of nucleosomes in chromatin: specific or statistical? Nature 292 (1981) 579–580.
[4] H.D. Ou, S. Phan, T.J. Deerinck, A. Thor, M.H. Ellisman, C.C. O'Shea, Chromemt: visualizing 3d chromatin structure and compaction in interphase and mitotic cells, Science 357.
[5] J. Fraser, I. Williamson, W.A. Bickmore, J. Dostie, An overview of genome organization and how we got there: from fish to hi-c, Microbiol. Mol. Biol. Rev. 79 (2015) 347–372.
[6] A.S. Hansen, C. Cattoglio, X. Darzacq, R. Tjian, Recent evidence that tads and chromatin loops are dynamic structures, Nucleus 9 (2018) 20–32.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

[7] K. Baumann, A vision of 3D chromatin organization, Nat. Rev., Mol. Cell Biol. 18 (2017) 532.
[8] G. Felsenfeld, M. Groudine, Controlling the double helix, Nature 421 (2003) 448–453.
[9] M. Ricci, C. Manzo, M.F. García-Parajo, M. Lakadamyali, M. Cosma, Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo, Cell 160 (2015) 1145–1158.
[10] J.C. Hansen, Human mitotic chromosome structure: what happened to the 30-nm fibre? EMBO J. 31 (2012) 1621–1623.
[11] D.J. Tremethick, Higher-order structures of chromatin: the elusive 30 nm fiber, Cell 128 (2007) 651–654.
[12] K.J. Meaburn, T. Misteli, Chromosome territories, Nature 445.
[13] S.V. Razin, A.A. Gavrilov, Chromatin without the 30-nm fiber: constrained disorder instead of hierarchical folding, Epigenetics 9 (5) (2014) 653–657.
[14] S.S. Rao, S.-C. Huang, B.G.S. Hilaire, J.M. Engreitz, E.M. Perez, K.-R. Kieffer-Kwon, A.L. Sanborn, S.E. Johnstone, G.D. Bascom, I.D. Bochkov, X. Huang, M.S. Shamim, J. Shin, D. Turner, Z. Ye, A.D. Omer, J.T. Robinson, T. Schlick, B.E. Bernstein, R. Casellas, E.S. Lander, E.L. Aiden, Cohesin loss eliminates all loop domains, Cell 171 (2017) 305–320, e24.
[15] W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N.A. Fonseca, W. Huber, C.H. Haering, L. Mirny, F. Spitz, Two independent modes of chromatin organization revealed by cohesin removal, Nature 551 (2017) 51–56.
[16] R.C. Allshire, H.D. Madhani, Ten principles of heterochromatin formation and function, Nat. Rev., Mol. Cell Biol. 19 (2018) 229–244.
[17] M. Crochemore, L. Ilie, W. Rytter, Repetitions in strings: algorithms and combinatorics, Theoret. Comput. Sci. 410 (2009) 5227–5235.
[18] K.G. Lim, C.K. Kwoh, L.Y. Hsu, A. Wirawan, Brief. Bioinform. 14 (2013) 67–81.
[19] A.J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications, Cell Res. 21 (2011) 381–395.
[20] N. Krietenstein, M. Wal, S. Watanabe, B. Park, C.L. Peterson, B.F. Pugh, P. Korber, Genomic nucleosome organization reconstituted with pure proteins, Cell 167 (2016) 709–721.
[21] M. Radman-Livaja, O. Rando, Nucleosome positioning: how is it established, and why does it matter? Dev. Biol. 339 (2) (2010) 258–266.
[22] C. Jiang, B. Pugh, Nucleosome positioning and gene regulation: advances through genomics, Nat. Genet. 10 (2010) 161–172.
[23] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, J. Moore, J. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[24] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, E.M. LeProust, T.R. Hughes, J.D. Lieb, J. Widom, E. Segal, The DNA-encoded nucleosome organization of a eukaryotic genome, Nature 458 (2009) 362–366.
[25] R. Kornberg, L. Stryer, Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism, Nucleic Acids Res. 16 (1988) 6677–6690.
[26] V. Charoensawan, S. Janga, M. Bulyk, M. Babu, S. Teichmann, DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes, Mol. Cell 47 (2012) 183–192.
[27] G. Locke, D. Haberman, S.M. Johnson, A.V. Morozov, Global remodeling of nucleosome positions in C. *elegans*, BMC Genomics 14 (2013) 284.
[28] E. Segal, J. Widom, What controls nucleosome positions? Trends Genet. 746 (2009) 1–9.
[29] E. Segal, J. Widom, Poly(dA:dT) tracts: major determinants of nucleosome organization, Curr. Opin. Struck. Biol. 19 (2009) 65–71.
[30] Y. Lorch, B. Maier-Davis, R.D. Kornberg, Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions, Genes Dev. 28 (2014) 2492–2497.
[31] R. Blossey, H. Schiessel, The latest twists in chromatin remodeling, Biophys. J. (2018) 2255–2261.
[32] T. Mavrich, I. Ioshikhes, B. Venters, C. Jiang, L. Tomsho, J. Qi, S. Schuster, I. Albert, B.F. Pugh, A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome, Genome Research.
[33] V.B. Teif, Nucleosome positioning: resources and tools online, Brief. Bioinform. 17 (2016) 745–757.
[34] M. Heinig, M. Colomé-Tatché, A. Taudt, C. Rintisch, S. Schafer, M. Pravenec, N. Hubner, M. Vingron, F. Johannes, histonehmm: Differential analysis of histone modifications with broad genomic footprints, BMC Bioinform. 16 (2015) 60.
[35] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (2015) 321–332.
[36] J. Zhong, T. Wasson, A.J. Hartemink, Learning protein–dna interaction landscapes by integrating experimental data through computational models, Bioinformatics 30 (2014) 2868–2874.
[37] W. Möbius, U. Gerland, Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites, PLoS Comput. Biol. 6 (2010) e891.
[38] D.E. Schones, et al., Dynamic regulation of nucleosome positioning in the human genome, Cell 132 (2008) 887–898.
[39] V.B. Teif, Nucleosome positioning: resources and tools online, Brief. Bioinform. 17 (2016) 745–757.
[40] A. Valouev, J. Ichikawa, T. Thaisan, J. Stuart, R. Swati, H. Peckham, K. Zeng, J. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, S.M. Johnson, A high-resolution, nucleosome position map of C. *elegans* reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.
[41] F. Utro, V. Di Benedetto, D.F. Corona, R. Giancarlo, The intrinsic combinatorial organization and information theoretic content of a sequence are correlated to the DNA encoded nucleosome organization of eukaryotic genomes, Bioinformatics 32 (2016) 835–842.
[42] E. Trifonov, Making sense of the human genome, in: Human Genome Initiative and DNA Recombination, Vol. 1 of Structure and Methods, Proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics, 1990, pp. 68–78.
[43] A. De Luca, S. Varricchio, Finiteness and Regularity in Semigroups and Formal Languages, Monogr. Theoret. Comput. Sci. EATCS Ser., Springer, Heidelberg, Germany, 1999.
[44] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, New York City, NY, USA, 1991.
[45] M. Li, P.M.B. Vitányi, An Introduction to Kolmogorov Complexity and Its Application, Springer-Verlag, New York City, NY, USA, 1997.
[46] R. Giancarlo, D. Scaturro, F. Utro, Textual data compression in computational biology: a synopsis, Bioinformatics 25 (2009) 1575–1586.
[47] R. Giancarlo, D. Scaturro, F. Utro, Textual data compression in computational biology: algorithmic techniques, Soc. Sci. Comput. Rev. 6 (2012) 1–25.
[48] P. Ferragina, R. Giancarlo, G. Manzini, M. Sciortino, Boosting textual compression in optimal linear time, J. ACM 52 (2005) 688–713.
[49] T. Mavrich, C. Jiang, I. Ioshikhes, X. Li, B. Venters, S. Zanton, L. Tomsho, J. Qi, R. Glaser, S. Schuster, D. Gilmour, I. Albert, B. Pugh, Nucleosome organization in the *Drosophila* genome, Nature 453 (2008) 358–364.
[50] A. Doring, D. Weese, T. Rausch, K. Reinert, SeqAn an efficient, generic C++ library for sequence analysis, BMC Bioinform. 9 (2008) 11, http://www.biomedcentral.com/1471-2105/9/11.
[51] D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, New York City, NY, USA, 1997.
[52] R. Giancarlo, S. Rombo, F. Utro, Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies, Brief. Bioinform. 15 (2014) 390–406.
[53] M.D. Cao, T.I. Dix, L. Allison, C. Mears, A simple statistical algorithm for biological sequence compression, in: Proc. of the IEEE Data Compression Conference (DCC), IEEE Computer Society, 2007, pp. 43–52.
[54] I.H. Witten, R.M. Neal, J.G. Cleary, Arithmetic coding for data compression, Commun. ACM 30 (1987) 520–540.
[55] J.Y. Chen, S. Lonardi, Biological Data Mining, Chapman and Hall, 2009.
[56] F. Zambelli, G. Pesole, G. Pavesi, Motif discovery and transcription factor binding sites before and after the next-generation sequencing era, Brief. Bioinform. 14 (2013) 225–237.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047

[57] S.E. Rombo, F. Utro, R. Giancarlo, Basic Statistical Indices for SeqAn, Chapman & Hall/CRC Mathematical & Computational Biology.
[58] A. Apostolico, M.E. Bock, S. Lonardi, Monotony of surprise and large-scale quest for unusual words, J. Comput. Biol. 10 (2/3) (2003) 283–311.
[59] H. Peckham, R. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W. Noble, K. Struhl, Z. Weng, Nucleosome positioning signals in genomic dna, Genome Res. 17 (2007) 1170–1177.
[60] D. Tillo, T. Hughes, G+C content dominates intrinsic nucleosome occupancy, BMC Bioinform. 10 (2009) 442.
[61] R. Giancarlo, S.E. Rombo, F. Utro, Epigenomic k-mer dictionaries: shedding light on how sequence composition influences nucleosome positioning *in vivo*, Bioinformatics 31 (2015) 2939–2946.
[62] U. Ferraro Petrillo, G. Roscigno, G. Cattaneo, R. Giancarlo, Informational and linguistic analysis of large genomic sequence collections via efficient hadoop cluster algorithms, Bioinformatics (2018) bty018, https://doi.org/10.1093/bioinformatics/bty018.
[63] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[64] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.
[65] M. Triola, Elementary Statistics, 12th edition, Pearson, San Francisco, U.S.A., 2012.
[66] A. Gordon, Null models in cluster validation, in: W. Gaul, D. Pfeifer (Eds.), From Data to Knowledge, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, Heidelberg, 1996, pp. 32–44.
[67] R. Giancarlo, D. Scaturro, F. Utro, A tutorial on computational cluster analysis with applications to pattern discovery in microarray data, Math. Comput. Sci. 1 (4) (2008) 655–672.
[68] R. Giancarlo, F. Utro, Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis, Theoret. Comput. Sci. 428 (2012) 58–79.
[69] M. Ghodsi, B. Liu, M. Pop, DNACLUST: accurate and efficient clustering of phylogenetic marker genes, BMC Bioinform. 12 (2011) 271.
[70] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[71] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.
[72] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A.C. Thastrom, Y. Field, I.K. Moore, J.-P.Z. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[73] S. Liu, L. Zhang, H. Quan, H. Tian, L. Meng, L. Yang, H. Feng, Y. Q. Gao, From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective, bioRxiv, https://doi.org/10.1101/255174.

Please cite this article in press as: R. Giancarlo et al., DNA combinatorial messages and Epigenomics: The case of chromatin organization and nucleosome occupancy in eukaryotic genomes, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.047