Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC

Amanda Hughes
University of Massachusetts Medical School

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/oapubs

Part of the Biochemistry, Biophysics, and Structural Biology Commons, Genetics and Genomics Commons, and the Medicine and Health Sciences Commons

Repository Citation
Hughes A, Rando OJ. (2009). Chromatin 'programming' by sequence - is there more to the nucleosome code than %GC. Open Access Articles. https://doi.org/10.1186/jbiol207. Retrieved from https://escholarship.umassmed.edu/oapubs/2146

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Open Access Articles by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Minireview

Chromatin ‘programming’ by sequence - is there more to the nucleosome code than %GC?

Amanda Hughes and Oliver J Rando

Address: Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street No. 903, Worcester, MA 01605, USA.

Correspondence: Oliver Rando. Email: oliver.rando@umassmed.edu

Abstract
The role of genomic sequence in directing the packaging of eukaryotic genomes into chromatin has been the subject of considerable recent debate. A new paper from Tillo and Hughes shows that the intrinsic thermodynamic preference of a given sequence in the yeast genome for the histone octamer can largely be captured with a simple model, and in fact is mostly explained by %GC. Thus, the rules for predicting nucleosome occupancy from genomic sequence are much less complicated than has been claimed.

See research article http://www.biomedcentral.com/1471-2105/10/442.

Packaging of eukaryotic DNA into nucleosomes has profound effects on DNA-templated processes. The 147 bp of DNA wrapped around the histone octamer is generally believed to be less accessible to DNA-binding proteins than is the DNA between nucleosomes. The positioning of nucleosomes relative to underlying sequences therefore has considerable implications for the regulation of gene expression, and understanding where nucleosomes are located and the rules underlying nucleosome positioning are key questions in understanding transcriptional control.

The recent revolution in genomics technologies has made genome-wide mapping of nucleosome positions possible in organisms ranging from budding yeast to humans. These genome-wide maps provide us with a multitude of hypotheses regarding the role of nucleosome positioning in gene regulation (reviewed in [1-3]). Perhaps one of the biggest surprises from even the earliest of these mapping efforts (in *Saccharomyces cerevisiae*) was the observation that the majority of nucleosomes are ‘well positioned’, that is, that nucleosomes occupy the same position (in many cases, to within mapping precision) in the majority of cells in a mixed population in the mid-log phase of growth (that is, actively growing unsynchronized yeast). This was a surprise to many investigators for many reasons, not least because the a priori expectation for a general packaging-protein complex would include a lack of sequence specificity. Furthermore, yeast promoters turned out to look very similar to one another, with a nucleosome-depleted 'nucleosome-free region' (NFR) observed at the majority of yeast promoters. This unanticipated level of order then raises the question of what underlies the remarkably consistent chromatin packaging in cell populations. Work recently published in *BMC Bioinformatics* by Tillo and Hughes [4] provides one surprisingly simple answer to this question, suggesting that the rules for predicting nucleosome occupancy from genomic sequence may be much less complicated than had been widely supposed.

The positioning of nucleosomes in vivo

Some properties of strongly pro- and antinucleosomal DNA sequences had already been elaborated in the pre-genomic era, but given the limited DNA sequencing capacity available, the extent to which genomic sequence programmed chromatin structure in vivo was unknown. Nucleosome positioning at any given locus can be ascribed to either local cis sequence cues or trans-acting protein factors (or, of course, both). Chromatin-remodeling complexes can move or evict nucleosomes, providing the canonical examples of trans-acting factors [5]. Conversely, it has been known for decades that there is at least some variation between DNA sequences in their affinity for the histone octamer [6]. The basic insight that led to this realization came originally from the observation that some DNA sequences were more or less flexible. Because DNA is sharply bent around the histone octamer, stiff sequences should be less favorable for nucleosomal incorporation, whereas flexible sequences or intrinsically curved sequences would be more favorable sites for octamer placement. In early studies, polyA sequences were shown to be intrinsically stiff, apparently owing to systems of 'bifurcated' hydrogen bonds between a given A and two Ts on the opposite strand. Conversely, because AT dinucleotides potentially introduce a kink in DNA, spacing of AT dinucleotides every 10 bp would be expected to result in DNA with a consistent curvature, reducing the free-energy cost of bending these sequences and resulting in more thermodynamically stable nucleosomes.

Since these early observations, the extent to which genomic sequence directs chromatin structure through intrinsic
preferences has been an active area of investigation. One approach involved investigations in vivo - seminal studies from the Struhl group in S. cerevisiae showed that nucleosome depletion at the HIS3 promoter could be enhanced or diminished by adding or removing polyA sequences [7]. However, in vivo studies are subject to the criticism that it is nearly impossible to exclude possible effects of an unknown trans-acting polyA-binding protein, although in S. cerevisiae this appears to be unlikely to account for the observations.

A general way to demonstrate that a given sequence intrinsically favors or disfavors nucleosome incorporation is to carry out in vitro nucleosome-reconstitution assays using nothing but histones, DNA and buffer. For instance, Struhl and colleagues showed that nucleosome depletion at the HIS3 promoter can be recapitulated in vitro [8], while Korber and colleagues showed that the PHO5 promoter can only be assembled into its in vivo packaging in the presence of yeast extract [9]. Wide-ranging studies from various groups over decades has provided a great deal of insight into the rules underlying histone-DNA interactions. For example, selections for tight-binding sequences provided the chromatin community with the best-defined ‘pronucleosomal’ sequence, the ‘Widom601’ sequence (identified by Jon Widom and colleagues), which has been used in countless in vitro studies [3].

The subsequent sequencing of numerous genomes and the advent of genomic nucleosome maps provided fodder for a range of computational studies (reviewed in [1-3]). Initial studies focused on pronucleosomal sequences with a 10-nucleotide periodicity of AT, AA, or TT dinucleotides. Two early studies agreed that such sequences were enriched at the +1 nucleosome position, but these studies did not capture the dominant feature of yeast promoters - nucleosome depletion at the so-called nucleosome-free region. Subsequent studies from many groups improved on these models by systematically incorporating antinucleosomal sequences (such as polyA and others) that are prevalent at yeast promoters and appear to be a major determinant of nucleosome-free regions in vivo.

**In vitro reconstitution studies reveal ‘programmed’ nucleosome-free regions**

All of the studies noted above focused on predicting in vivo nucleosome positions, since in vitro reconstitution data were sparse. More recently, two groups have carried out genome-wide experimental studies of intrinsic nucleosome-binding preferences [10,11]. These studies differed in their conclusions, but the data are quite similar. In essence, in vitro reconstitution of yeast genomic DNA into nucleosomes captures nucleosome depletion at yeast promoters, but little else (Figure 1). The periodic spacing of AA/AT/TT dinucleotides that is statistically enriched at the in vivo +1 position does not appear to play a general role in positioning the +1 nucleosome, and more probably ‘fines-tunes’ rotational positioning of nucleosomes (that is, positioning to ±1 nucleotide after large-scale ±5 nucleotide positioning has been established by other means [11]).

Kaplan et al. [10] argue on the basis of a high (around 0.74) correlation coefficient between the in vitro and in vivo datasets that in vitro reconstitutions globally capture in vivo chromatin architecture. However, as Stein et al. [12] recently showed, the use of correlation coefficients is misleading because they are subject to the ‘influential point effect’ - in other words, outlying points drive correlations (even if the bulk of the data are uncorrelated), and in the case of chromatin structure these outlying points correspond to the dramatic nucleosome depletion at promoters. Indeed, Zhang et al. [11] showed very poor correspondence between in vivo nucleosome positioning and in vitro reconstitution data. Thus, we believe that all extant data support a view in which very little nucleosome positioning information is intrinsically encoded but that the yeast genome does program nucleosome depletion at promoters via antinucleosomal sequences. Below, we will
discuss sequence models that attempt to recapitulate the in vitr o reconstitution results, but readers should bear in mind that nucleosome occupancy rather than nucleosome positioning is being addressed when correlation coefficients are being used as the summary statistic.

**Trimming the fat from computational models**

In addition to establishing that intrinsic ‘programming’ of chromatin architecture is largely limited to nucleosome depletion at promoters, in vitr o reconstitution datasets also enable more direct testing of computational models of intrinsic sequence preferences for nucleosomes. For instance, Kaplan et al. [10] generated a model to predict the intrinsic preference of a given 147-bp sequence for nucleosomes. This model consists of a position-independent component (that is, a component that does not depend on location within the 147-bp sequence), and a position-dependent component. The position-independent component was based on measured occupancy of all 5mer (5-nucleotide) sequences - some 5mers, such as AAAAA, are rarely found in nucleosomes and thus a sequence carrying AAAAA would be weighted unfavorably for nucleosome formation. The position-dependent term builds on the dinucleotide periodicity noted above: for each position relative to the dyad axis, the frequency of each dinucleotide was calculated from reconstitution data, and then 147-bp sequences were scored for their match to this distribution. Predictions of this model correlated very well (0.89) with in vitr o reconstitution data, suggesting that the majority of the affinity of a given sequence for the histone octamer is predictable from sequence (again, note that the use of correlation coefficients emphasizes outlier sequences, such as the polyAs found at promoters).

The new work by Tillo and Hughes [11] now extends the search for sequence rules underlying nucleosome occupancy. Given the large number (more than 2,000) of parameters included in the Kaplan model (all 5mers, plus dinucleotide frequencies across 127 bp), Tillo and Hughes asked whether a simpler model might capture most of the occupancy information encoded in DNA sequence. They used a linear regression algorithm called Lasso to identify features that predict nucleosome occupancy in the Kaplan dataset. Specifically, after selecting a large number of candidate features (straightforward candidate features such as %GC content, not complex eigenvectors such as generated by principal component analysis), Lasso creates a linear combination model with an emphasis on setting as many coefficients to zero as possible. The resulting model(s) had very few parameters, with the model selected for study having only 14 features.

The resulting sequence model captures in vitr o nucleosome occupancy data nearly as well (R = 0.86) as the Kaplan model (R = 0.89), indicating that a small number of sequence features describes most of the variation in nucleosome occupancy in in vitr o reconstitutions. Close examination of the most important features of this model indicates that %GC and polyA runs are the two dominant factors, with a simple model using just these features exhibiting a correlation of 0.72 with in vitr o data. Indeed, a model based on %GC alone showed a correlation of 0.71 with the in vitr o data. Much of this is likely to be a consequence of the fact that many of the other 13 parameters are correlated with %GC - AAAAA is obviously unlikely in high-GC sequences. Furthermore, many features of DNA three-dimen sional structure (the authors specifically note ‘propeller twist’ and ‘slide’) are also correlated with %GC, and thus GC content seems to provide a single feature that captures many related structural characteristics that are important for nucleosome stability. Additional features in the model (AAAA, propeller twist, and so on) are then proposed to indicate features that are important for nucleosome formation but are not entirely captured by GC content. It is important to consider that there may also be a confounding effect of genome structure - yeast promoters are AT-rich and nucleosome-depleted, so %GC will naturally correlate with nucleosome depletion whether it is a cause or a consequence - but the authors also compare their model with data from synthetic DNA reconstitution data and still obtain significant correlations with the in vitr o data.

These results have a number of important implications for thinking about how chromatin structure is ‘programmed’. First, as there are no terms for dinucleotide periodicity in the Lasso model, these results support the finding of many groups arguing that nucleosome exclusion by polyA and related sequences is the dominant feature in nucleosome reconstitution assays. Second, the lack of support for ‘pronucleosomal’ sequences as major positioning cues in the reconstitution data re-emphasizes the status of statistical positioning as the best hypothesis to explain why chromatin is so well ordered in vivo . Third, because the Tillo and Hughes model also performs reasonably well on nucleosome-mapping data from Caenorhabditis elegans, it may prove portable for analysis of genomes other than yeast. Finally, these results have important implications for genome structure and evolution, as GC content varies between organisms and across genomes (CpG islands being a prominent example).

The recent lively interest in the idea of a ‘nucleosome code’ that might program the packaging of the genome thus seems somewhat excessive. Tillo and Hughes help clear the air to some extent, showing that simple models very effectively capture the majority of the behavior of in vitr o nucleosome reconstitution experiments. Programming nucleosome depletion with AT-rich sequences at promoters is confirmed as a key regulatory strategy in budding yeast and perhaps C. elegans. These insights may help guide questions about the evolution of chromatin packaging at specific loci, and about the regulatory strategies available...
to promoters with large nucleosome-free regions ‘programmed’ in cis.

Acknowledgements
We thank M Radman-Livaja and N Friedman for critical comments on this minireview. OJR is supported in part by a Career Award in the Biomedical Sciences from the Burroughs Wellcome Fund, and by grants from NIGMS and HFSP.

References
1. Jiang C, Pugh BF: Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 2009, 10:161-172.
2. Radman-Livaja M, Rando OJ: Nucleosome positioning: How is it established, and why does it matter? Dev Biol 2009 doi:10.1016/j.ydbio.2009.06.012.
3. Segal E, Widom J: What controls nucleosome positions? Trends Genet 2009, 25:335-343.
4. Tillo D, Hughes TR: G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 2009, 10:442.
5. Clapier CR, Cairns BR: The biology of chromatin remodeling complexes. Annu Rev Biochem 2009, 78:273-304.
6. Drew HR, Travers AA: DNA bending and its relation to nucleosome positioning. J Mol Biol 1985, 186:773-790.
7. Struhl K: Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci USA 1985, 82:8419-8423.
8. Sekinger EA, Moqtaderi Z, Struhl K: Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol Cell 2005, 18:735-748.
9. Korber P, Horz W: In vitro assembly of the characteristic chromatin organization at the yeast PHOS promoter by a replication-independent extract system. J Biol Chem 2004, 279:35113-35120.
10. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E: The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 2009, 458:362-366.
11. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K: Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 2009, 16:847-852.
12. Stein A, Takasuka TE, Collings CK: Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? Nucleic Acids Res 2009, doi:10.1093/nar/gkp1043.
13. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N: High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res 2009, doi: 10.1101/gr.098509.109.

Published: 23 December 2009
doi:10.1186/jbiol207
© 2009 BioMed Central Ltd