Predicting nucleosome positions from DNA sequence

Kristof Redei

November 26, 2007

Abstract

DNA carries genetic information needed to construct cell components like proteins and RNA. The encoding through which certain sequences of DNA generate specific amino acids is well understood. Recent studies have shown that in addition to this coding scheme, DNA also influences a part of its own physical arrangement. By constructing a model of the interaction between DNA and nucleosomes, the compacted units of 147 base pairs wrapped around a histone octamer, it is possible to predict aspects of nucleosome organization from sequence information alone. This organization, in turn, has a significant effect on cellular processes such as transcription and gene expression.

1 The nucleosome

In eukaryotes, the nucleosome is the fundamental repeating unit of chromatin, the complex of DNA and proteins that forms the chromosome. A nucleosome is made up of 147 base pairs coiled around eight histone molecules. Histones are proteins that were discovered as early as the 19th century and compose, by mass, roughly an equal amount of the nucleus as DNA; thus they were once thought of as the main carriers of genetic information. [7] Five types of histones, whose amino acid sequence is mostly invariant across species, can be distinguished. Four of them (H2A, H2B, H3, and H4) make up the octamer around which DNA coils, while the fifth (H1) attaches to the outside of the nucleosome. Nucleosomes are connected by histone-free linker DNA whose length is variable, usually between 10-50 base pairs. Hence, 75-90% of the DNA of a eukaryotic organism is part of a nucleosome. [13] This structure is often illustrated by comparing it to a series of "beads on a string."

The ability of histone octamers to form nucleosomes with DNA is significant for several reasons. First, wrapping DNA in histones makes it compact enough to fit into the cell nucleus, whose diameter is on the scale of a few micrometers. Since the length of an uncoiled DNA molecule can be several degrees of magnitude larger, nucleosomes are essential in the formation of tightly packaged chromatin.

Second, and more significantly from the point of view of this paper, the presence of a nucleosome affects access to the DNA that it wraps. For example, the expression of genes at nucleosome-bound sites can be repressed since histones prevent RNA from binding to and transcribing them. One early discovery of this process was made by Han and Grunstein, who induced nucleosome depletion by repressing histone synthesis and observed an increase in transcription initiation in yeast. [4] This means that nucleosomes can be used to occlude sites that RNA might recognize but that have no function, since a large genome might contain many sites whose sequence is coincidentally similar to a useful functional site. Conversely, their absence or removal can expose these sites. [12]

Moreover, the positions of nucleosomes are not static and unchanging. Nucleosome occupancy can be regulated by a wide variety of chromatin remodeling complexes that organize positioning and move histones between different locations. [1] This may be useful when several steps are required to form a complex of multiple proteins. [12] In addition, it is not only the location of histones that can change, but also their composition. For example, acetylation of the tails of the histones reduces the repression of transcription. [7]

2 The influence of sequence

Nucleosomes' positioning thus has important biological effects. Remodeling complexes, as noted above, are one mechanism through which the cell regulates their placement. However, they are not the only factor in determining this distribution. A number of recent studies have shown that DNA sequence affects where nucleosomes are formed.

Certain DNA sequences are more likely to be bound to nucleosomes due to their intrinsic physical capacity for the tight bending that is required in order to coil around the histone octamer. [3] That this is the case has been known for at least two decades. Travers et al. [14] showed in 1987, using statistical analysis of 177 DNA molecules from chicken erythrocytes, that certain di- and trinucleotides within nucleosome-bound sequences exhibit well-defined periodicities based on the way specific bases influence the molecule's physical structure. These motifs, AA and TT with a period of about 10 bases, and GC with the same period but about 5 bases out of phase with AA/TT, mainly affect the rotational position of the helix relative to the histone. As for the translational orientation, i.e. the position in the sequence at which nucleosomes will form, Kunkel and Martinson [8] discovered that poly(dA)-poly(dT) sequences are disfavored by nucleosomes and cannot appear in them at all above certain lengths. Nelson et al. [10] showed that this is because such sequences form a helical structure that is too rigid for the bending needed in order to be included in a nucleosome.

Determining the extent to which sequence affinities affect nucleosome positioning requires accurate measurement of the latter. A recent study [17] used a tiled DNA microarray method to distinguish nucleosome-bound and linker DNA in the genome of budding yeast. A hidden Markov model (HMM) using the measured hybridization values as observable states and "nucleosomal" and "linker" as hidden states was used to generate data that corresponded well with published nucleosome positions. This allowed the identification of nucleosome positions across almost half a million bases. The large amount of information allowed more robust conclusions about the correlations between sequence and nucleosome position. Specifically, correlations between the locations of nucleosomes and those of functional binding and transcription start sites were discovered.

The study shed new light on the link between nucleosome positions and promoter sequences that are bound by a large number of transcription factors. Since such sites must be frequently accessed by RNA, they would presumably be more likely to be nucleosome-depleted. The data showed this to be the case as 87% of these sites were, in fact, depleted. The biological importance of transcription factor motifs causes them to be highly conserved, thus the conservation of nucleosome-free regions (NFRs) between related yeast species was examined. It was found that, compared to the average conservation of intergenic, non-coding DNA, NFRs were more highly conserved than sequences found in nucleosomes. These areas included not only transcription factor motifs, but long stretches of adenosine and thymidine, which had been shown in earlier studies like [8] to be difficult for histones to bind to because of their low capacity for bending.

These results regarding conservation led Yuan et al. to suggest a rough model of how sequence affects nucleosome positions. Over two-thirds of DNA were found to appear in nucleosomes that are well-positioned, meaning that their location in the sequence is well-defined, as opposed to "delocalized" nucleosomes, whose position may vary rapidly along a longer string due to thermodynamic preferences. The latter also tend to appear farther from areas that are consistently nucleosome-free. Their idea is thus that NFRs have consistent locations on the sequence due to the causes above; nucleosome positions, in turn, are constrained by NFR locations.

3 Prediction from sequence

Thanks to the experimental data collected by Yuan et al, several subsequent attempts have been made to approach the sequence-nucleosome connection from a computational perspective. Given the empirical information on where nucleosomes appear in the genome, probabilistic models can be designed that attempt to predict their locations. The predictions, in turn, can be evaluated by comparing them to experimentally measured positions.

3.1 Comparative genomics: sequence influences its regulation

One of the first of these attempts [5] involved the use of comparative genomics to reduce the noise present in raw sequence data. In an earlier study [6], a probabilistic model was derived through multiple sequence alignment of 204 DNA fragments known to be nucleosome-bound. Building on the results mentioned earlier in which the significance of the dinucleotides AA and TT in nucleosome positioning was recognized, the model attempted to characterize the frequency distributions of these two dinucleotides across sequences known to form nucleosomes. Computing the correlation with this model of each location from -1000 to +800 bases relative to the start codon in yeast genes yields a profile that is meant to reveal the locations of nucleosome positioning sequences (NPS) and areas that are unlikely to be nucleosome-bound. Since individual sequences tend to contain too much noise to achieve a clear result, predictions were averaged across groups of genes to strengthen the effect of relevant patterns. Using various criteria for grouping can thus provide insight about differences in positioning for distinct classes of genes, hence this method is dubbed the "comparative genomics" approach.

Comparison of the results from this computational approach to the experimental nucleosome map showed strong similarity, with correlations to the NPS pattern from the model corresponding closely to mapped locations of nucleosomes, and negative correlations to nucleosome-free regions. The close correspondence of the results from the two methods increases confidence in the validity of both. The first approach to grouping genes distinguished two categories: the top 15% of all genes either strongly positively or negatively regulated by histone tail modifications. Differences between these two groups were found only at the sites of assembly of transcription machinery, indicating that promoters of these two distinct types have different positioning properties.

Since genes with positive nucleosome regulation tend to contain a DNA sequence called a TATA box in their promoters, these genes were then examined separately. One location in which NFRs were mapped experimentally but not predicted by the model was the area directly around the TATA boxes. It was hypothesized that averaging correlations over all TATA-containing genes in the data set may have obscured relevant information if some of the TATA-containing genes have strong positive NPS correlations in that area while others are negatively correlated; the two signals may have cancelled each other out.

Indeed, further refining the grouping by averaging instead over orthologous locations in six different yeast species produced much greater correlations that corresponded better to the experimental data. These sequences were subsequently clustered into three distinct groups with two of them displaying robustly opposite nucleosome affinities at the region in question, showing that examining groups of orthologous genes separately instead of an aggregate of all promoters can reveal additional, useful information. It was found that the cluster whose TATA box tended to be nucleosome-wrapped contained a large proportion of genes that are highly regulated by chromatin remodeling factors. Thus, control of transcription is made possible by wrapping the gene in a nucleosome that can be repositioned by these factors when needed. On the whole, then, the results of Ioshikhes et al show that computational prediction of nucleosome positions can lead to reasonably accurate results, and the use of comparative genomics can elucidate differences in how positions arise in distinct types of genes. More broadly, not only can dinucleotides be shown to influence nucleosome positions, but distinct gene types can be shown to correlate with distinct placement patterns.

3.2 Sequence: a code for position, organization, and transcription?

A slightly different probabilistic model was used by Segal et al [13]. While also computing dinucleotide probability distributions, they made no a priori assumption about which of them are correlated with nucleosome occupancy. First, a collection of 199 nucleosome-bound DNA sequences was experimentally generated. A model similar to a PSSM was then generated based on these sequences. The score computed at each location of a sequence represents the predicted average nucleosome occupancy at that base.

The model differs from a standard PSSM in several ways. The collection is center-aligned and the reverse complement of each sequence is added. This is done in order to account for symmetry in the structure of nucleosomes. While a regular PSSM expresses mononucleotide probability distributions (i.e. the chance of seeing a given base at a given position), dinucleotide distributions are computed in the DNA-nucleosome model. This is done because of their link to physical structure and bending capacity, as discussed above. Finally, the distribution is smoothed by estimating the probability for each dinucleotide at each position i from the sum of counts at $i-1$, i, and $i+1$. Experimental results indicate that a 1-base change in nucleosome spacing can occur with small cost.

The model and the method used to generate it were validated in several ways. Using different training input - nucleosome-bound sequence collections from chicken, mouse, and random synthetic DNA - produced similar profiles. Experiments were carried out with sequences modified to increase or decrease their agreement with the motifs in the model, and measured nucleosome affinity changed as predicted (better agreement produced higher affinity). The genome of yeast also contained a larger proportion of sequences with high predicted affinity than would occur by chance.

Standard dynamic programming algorithms were then used to determine the most likely arrangement of nucleosomes for the entire sequence, resulting in a map of nucleosome organization for the yeast genome. Comparison to *in vivo* measured positions for various locations mapped in other studies yielded a promising result: 54% of the predicted nucleosome positions were within 35 base pairs of those found in the experiments, compared to the 39 +/- 1% expected by chance. When compared to the positions mapped by Yuan et al., similar results were found (45% versus 32 +/- 1%). The model constructed from chicken DNA, when applied to yeast sequence, also showed strong correlation, indicating a degree of species-independence of the sequence-nucleosome interaction.

To support the notion that sequence is a strong determinant of nucleosome position, locations were experimentally mapped *in vitro* for yeast DNA and these were compared to the aforementioned maps from other studies. Again, a strong overlap was found. Together, these results lead the authors to claim that about half of the nucleosome positions in the genome are attributable to the sequence in and of itself.

The data allowed some conclusions regarding the interaction between sequence and higher-order properties of nucleosome positions. The distribution of pairwise distances between nucleosome centers showed strong regularity, with multiples of the average distance of 177 bp between centers occurring far more frequently than expected by chance. Thus sequence seems to influence the structure of chromatin indirectly through its effect on nucleosome placement.

Similarly to Ioshikhes et al's study, predicted occupancy rates for specific areas of the genome were related to their rate of expression and to their function. Specifically, occupancy was predicted to be low for highly expressed genes such as ribosomal RNA and transfer RNA genes. On the other hand, genes with variable expression (e.g. genes repressed under stress) tended to encode high occupancy, leading to the conclusion that other factors are responsible for reducing repression when this is needed. Genes were also grouped by function to determine whether certain gene types are predicted to have high or low occupancy; this turned out to be the case for many of the groups. This explains the success of the approach taken by Ioshikhes et al.

Segal et al. also put forth a hypothesis on occupancy at transcription factor binding sites: since such sites may occur by chance at random, functionally irrelevant locations in the genome, these locations may encode higher nucleosome occupancy to deter transcription factors and direct them to the appropriate location. Predicted occupancy was indeed found to be higher at such sites than at useful ones. As for the transcriptional start sites examined by Ioshikhes et al, these sites and the TATA boxes they contain were found most likely to occur in unoccupied linker DNA. This result is notably contradicted by the comparative genomics study, which concluded that TATA boxes can be positively or negatively correlated with nucleosomes, depending on the specific gene promoter they belong to.

3.3 A whole-genome nucleosome map: structure is key

The most recent study on nucleosome position mapping and prediction provides a wealth of new data that provides an opportunity to test various models of the effects of sequence. [9] Lee et al's was the first study to provide a nucleosome occupancy map at high resolution for the whole of the yeast genome.

The study confirms earlier results regarding higher nucleosome occupancy at promoters of less expressed genes. The first new insight the analysis of the data provided was that aligning genes by their promoters' transcription start site (TSS) provides a more pronounced occupancy profile than aligning them according to their start codon, as was done in earlier studies. [5] Specifically, a nucleosome-depleted region is found just upstream of the TSS. This is not surprising considering the interactions between different promoter types and occupancy profiles (in particular, TATA-containing and TATA-less promoters) discovered earlier.

Building on this result, genes were then clustered based on differences in their nucleosome profiles to determine whether the latter were correlated with other properties. Contrary to the three-cluster result obtained by Ioshikhes et al, where distinct positioning profiles for TATA-containing genes were found to correlate most strongly with difference in regulation by chromatin remodelling complexes, here, k-means clustering with different values of k showed four roubstly distinct clusters. Each of these was found to contain higher proportions of genes with different functional characteristics, reinforcing the link between positioning profile and function.

A method for prediction of positioning from sequence, different from both of the previously described models, was then evaluated. Since transcription start sites seemed to play an important role in positioning, the correlation between nucleosome occupancy and the presence of transcription factor binding sites (TFBSs) in a gene was examined. A strong correlation was found, and some specific localization of the TFBS about 80-100 base pairs upstream of the TSS was possible, which corresponded to part of the depleted region described earlier.

However, TFBS positions cannot explain the positioning patterns for the whole of the genome by themselves. Therefore a statistical algorithm (Lasso) was used to construct a linear model by choosing the most strongly correlated features from a wide range that may have an influence, such as di- and tetranucleotides, various structural features, and Z-DNA. Interestingly, dinucleotides, which were used for prediction by both [5] and [13], were not selected as informative features; instead, a few structural features (tip, tilt, propeller twist) showed the strongest correlation, with the AAAA tetranucleotide and a few TFBSs also highly ranked. As mentioned earlier, poly(dA) sequences had been shown earlier to create a rigid physical structure that disfavors nucleosomes. On the whole, then, the results of the study seem to show that structural features that affect DNA bending seem to have the greatest effect on nucleosome position.

4 Conclusions and future work

The significant influence of DNA sequence on nucleosome positioning has thus been established and will continue to be explored in detail. The predictive power of sequence is such that some have stated that the positions of nucleosomes are "encoded" in the genome. [13] [2] Others are more skeptical about the degree of prediction possible.

Technical limitations of current studies can be distinguished from theoretical limits on prediction from sequence due to limited correlation and the influence of other factors. The former include bounds on the resolution and size of nucleosome maps; the recent whole genome map created by Lee et al. will doubtless be helpful in this regard. One of the clear directions for future research is the use of their information to test subsequent models of sequence-nucleosome interaction. For example, since the linear model reveals that certain tetranucleotides are more strongly correlated with positioning that the dinucleotides that probabilistic models have hitherto used, it would be interesting to see how the models in [5] and [13] would fare if modified to consider distributions of four-base sequences instead of pairs.

A broader problem is just how strong a predictor of nucleosome positions DNA is, in and of itself. Even the most accurate model will fail to reveal helpful information if the correlation is not robust enough. Chromatin remodeling complexes have already been mentioned as non-genomic agents that affect the sites to which nucleosomes bind *in vivo*. The pertinence of modifications of the tails of histones has also been shown in a recent study. [16] While the results of the study by Segal et al. are striking, it may be too strong a statement to say, as they do, that genetic information explains about half of the nucleosome organization, considering that their predictions are 15% above those expected by chance, and all predictions that are within 35 base pairs of literature positions are considered matches.

There is also some disagreement as to the appropriateness of the term "encoding" when discussing DNA-nucleosome interaction. While on its face, it may seem like a trivial semantic disagreement, the term may be misleading since it suggests a relationship similar to, for example, that between DNA triplets and amino acids. The latter constitutes a code since its symbols are arbitrary and have no direct causal link to the result that is ultimately produced; other, functionally equivalent codes could have evolved. This is not the case for nucleosome positioning; here, the physical properties engendered by the sequences directly affect the degree to which nucleosomes are able to bind to them. [15] Ioshikhes et al. also note that positioning sequences cannot be considered to constitute a code. Finally, the results showing that sequence exerts its influence on nucleosome positions most of all through its effect on structural properties [9] further weaken the case that specific sequence motifs stand for or represent certain nucleosome arrangements.

Code or not, the study of the interaction between sequence and nucleosomes has produced important insights into the functioning of the genome. Since the patterns discovered so far seem to be fairly robust across species, examination of chromatin structure in humans [11] may reveal some of the same regularities. Such studies will be helpful in further elucidating the links between genetic information and the organization of nucleosomes.

References

[1] Bradley R. Cairns. Chromatin remodeling complexes: strength in diversity, precision through specialization. *Current Opinion in Genetics and Development*, 2005(15):185–190, 11 February 2005.

[2] Sevinc Ercan and Jason D. Lieb. New evidence that dna encodes its packaging. *Nature Genetics*, 38:1104–1105, October 2006.

[3] Hernan G. Garcia, Paul Grayson, Lin Han, Mandar Inamdar, Jane Kondev, Philip C. Nelson, Rob Phillips, Jonathan Widom, and Paul A. Wiggins. Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers*, 2007(2):115–130, 5 February 2007.

[4] Min Han and Michael Grunstein. Nucleosome loss activates yeast downstream promoters in vivo. *Cell*, 55:1137–1145, 1988.

[5] Ilya P. Ioshikhes, Istvan Albert, Sara J. Zanton, and B. Franklin Pugh. Nucleosome positions predicted through comparative genomics. *Nature Genetics*, 38(10):1210–1215, October 2006.

[6] Ilya P. Ioshikhes, Alex Bolshoy, Konstantin Derenshteyn, Mark Borodovsky, and Edward N. Trifonov. Nucleosome DNA sequence pattern revealed by multiple sequence alignment of experimentally mapped sequences. *Journal of Molecular Biology*, 1996(262):129–139, 1996.

[7] Roger D. Kornberg and Yahli Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98:285–294, 6 August 1999.

[8] Gary R. Kunkel and Harold G. Martinson. Nucleosomes will not form on double-stranded RNA or over poly(dA)-poly(dT) tracts in recombinant DNA. *Nucleic Acids Research*, 9(24):6869–6888, 1981.

[9] William Lee, Desiree Tillo, Nicolas Bray, Randall H. Morse, Ronald W. Davis, Timothy R. Hughes, and Corey Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, 39(10):1235–1244, October 2007.

[10] Hillary C.M. Nelson, John T. Finch, Bonaventura F. Luisi, and Aaron Klug. The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature*, 330:221–226, 19 November 1987.

[11] Faith Ozsolak, Jun S. Song, X. Shirley Liu, and David E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, 25:244–248, January 2007.

[12] Timothy J. Richmond. Genomics: Predictable packaging. *Nature*, 442:750–752, 17 August 2006.

[13] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thastrom, Yair Field, Irene K. Moore, Ji-Ping Z. Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, 17 August 2006.

[14] A.A. Travers and A. Klug. The Bending of DNA in Nucleosomes and Its Wider Implications. *Phil. Trans. R. Soc. Lond.*, (317):537–561, 1987.

[15] Bryan M. Turner. Defining an epigenetic code. *Nature Cell Biology*, 9(1):2–6, January 2007.

[16] Iestyn Whitehouse and Toshio Tsukiyama. Antagonistic forces that position nucleosomes in vivo. *Nature Structural and Molecular Biology*, 13(7):633–640, July 2006.

[17] Guo-Cheng Yuan, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. Genome-Scale Identification of Nucleosome Positions in S. cerevisiae. *Science*, 309(5734):626–630, 2005.