Soft power of non-consensus protein-DNA binding

Vladimir B Teif[1,*]

[1]School of Life Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK.

*E-mail for correspondence: vteif@essex.ac.uk

Keywords: transcription factor, protein-DNA binding, specific binding, non-specific binding, DNA sequence repeats

In this issue, Goldshtein *et al* (1) show that if gene promoters are extended with DNA sequences containing repeating nucleotide patterns without specific protein-binding motifs, it is possible to predict the resulting changes in gene expression from so-called non-consensus protein-DNA binding. The authors found that during embryonic stem cell (ESC) differentiation, transcription factor (TF) preferences for such simple nucleotide repeats undergo distinct changes. This suggests an intriguing possibility that non-consensus binding may help direct TFs to different sub-classes of binding sites in different cell types.

DNA-protein binding has been studied in great detail for about a century. Chromatin proteins usually have positively charged domains that are naturally attracted to the negatively charged DNA, whatever the nucleotide sequence is. Such binding is traditionally called non-specific (2). However, most biologically interesting processes depend on the DNA sequence – this is when sequence-specific binding comes into play. In many classical examples of DNA-protein binding, a TF recognises a single stretch of nucleotides on the DNA (3), so-called consensus binding motif and its variations. Usually, the strength of such binding is several orders of magnitude higher than that of non-specific binding, which led to the concept of discrete TF binding sites – a limited number of small genomic regions where a given TF can bind (as opposed to the rest of the genome where TF binding can still happen but can be neglected due to its weakness). The concept of discrete TF binding sites has been very useful in predicting combinatorial, cooperative TF binding for several decades. Not all TFs appeared to recognise a single motif, but the concept of discrete binding sites could still hold by allowing several motifs for a single TF. The discrete binding site concept can be further extended to take into account new “letters” of the DNA alphabet arising due to naturally occurring chemical modifications such as methylation. However, in recent years, with the arrival of advanced computational methods such as deep learning on one hand and affordable high-throughput experiments on the other, it is becoming increasingly clear that many important TF-DNA binding events occur in the intermediate regime between non-specific and discrete site-binding (4).

One possibility to describe such intermediate binding regime is to widen the definition of the classical concept of consensus motifs to include, in addition to motifs responsible for unique DNA structures recognised by a single protein, new motifs responsible for several generic classes of local shape of the DNA double helix (such as widening of the narrow groove, bending, etc) (5, 6). Another possibility is to characterise protein-DNA binding beyond sequence motifs, searching for repeating nucleotide patterns – this is the approach that Goldshtein *et al* took (1) (Figure 1).

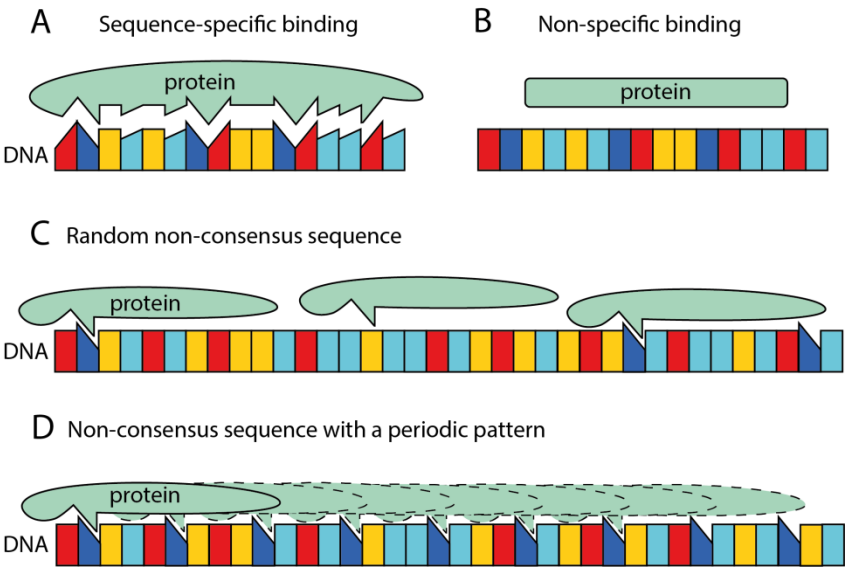

Figure 1. *Protein-DNA binding is not limited to sequence-specific (A) and non-specific (B), but can be also characterised by an intermediate regime of non-consensus binding which is non-specific for random DNA sequences (C) but is characterised by increased binding strength for DNA regions enriched with certain repeated nucleotide patterns (D). The DNA lattice units shown in different colours may correspond to individual base pairs or larger regions. Different geometric shapes of such units used on the figure do not imply that such shapes are actually visible in the DNA structure – these may also correspond to alterations of the DNA double helix stability or other slight perturbations of the energy landscape.*

Investigations of the role of genomic nucleotide periodicities in DNA-protein binding started about four decades ago, but were mainly in the context of nucleosome positioning (7). Nucleosomes cover most of the genome so the concepts of discrete binding site and a consensus motif naturally do not apply in this case, while the concept of nucleotide (or dinucleotide) oscillations appears quite handy. Nucleosome positioning is known to affect TF binding through competitive and cooperative interactions (8), but *direct* influence of DNA nucleotide periodicities on TF binding considered by Goldshtein *et al* (1) is a separate important effect.

In a series of recent publications Lukatsky and colleagues argued that stronger-than-random interaction of a DNA-binding protein with a genomic region containing simple nucleotide repeats has entropic nature: in this case it is statistically more probable for a protein to land on a region that contains its “favourite” repeat element (Figure 1D). They have performed *in vitro* experiments demonstrating that for natural genomic sequences the strength of this effect is quite significant and comparable to that of mutations in the specific TF motif (9). In the current paper, Goldshtein *et al* apply this approach to characterize TF-DNA binding in ESCs (1). For example, the authors showed that one of the key regulators of ESC development, c-Myc, possesses statistical preference for repetitive patterns of the type [CNNC] and [GNNG], where N stands for any nucleotide type. This computational prediction was verified in a plasmid reporter assay, introducing such repetitive patterns surrounding the consensus c-Myc binding sites at the promoter of the reporter gene. As predicted, this

resulted in higher gene expression in comparison with the case where flanking regions around c-Myc sites were composed of random DNA sequences.

The question about the physical nature of such non-consensus TF-DNA binding is still open. One possibility is that non-consensus DNA sequence repeats modulate the local stability of the DNA double helix. Goldshtein et al performed NMR measurements showing that DNA sequences with identical GC content but different DNA repeat symmetry types can indeed lead to different local DNA stabilities (1). Another possibility is that the non-consensus binding effects are due to slight changes of the local DNA shape, as in the models with DNA-shape motifs (5, 6) but in this case due to nucleotide oscillations in the absence of well-defined motifs. In a recent study that investigated 100 million random promoters, the effects on cis-regulatory logic associated with nucleotide changes outside TF binding motifs were mostly interpreted through the changes of DNA accessibility (10), but in light of the works mentioned above, this can be also explained by direct effects on non-consensus TF binding. Since genomes contain multitudes of simple repeats, such effects may have important roles in guiding differential TF binding during cell transitions, executing “soft power” on gene regulation beyond consensus motifs.

References

1. Goldshtein, M., M. Mellul, G. Deutch, M. Imashimizu, K. Takeuchi, E. Meshorer, O. Ram, and D. B. Lukatsky. 2020. Transcription factor binding in embryonic stem cells is constrained by DNA sequence repeat symmetry. Biophys J.
2. von Hippel, P. H., A. Revzin, C. A. Gross, and A. C. Wang. 1974. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. Proc Natl Acad Sci U S A 71(12):4808-4812.
3. Ptashne, M. 1967. Specific binding of the lambda phage repressor to lambda DNA. Nature 214(5085):232-234.
4. Inukai, S., K. H. Kock, and M. L. Bulyk. 2017. Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev 43:110-119.
5. Samee, M. A. H., B. G. Bruneau, and K. S. Pollard. 2019. A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs. Cell Syst 8(1):27-42 e26.
6. Zhou, T., N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordan, and R. Rohs. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci U S A 112(15):4654-4659.
7. Trifonov, E. N., and J. L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A 77(7):3816-3820.
8. Teif, V. B., F. Erdel, D. A. Beshnova, Y. Vainshtein, J. P. Mallm, and K. Rippe. 2013. Taking into account nucleosomes for predicting gene expression. Methods 62:26-38.
9. Afek, A., J. L. Schipper, J. Horton, R. Gordan, and D. B. Lukatsky. 2014. Protein-DNA binding in the absence of specific base-pair recognition. Proc Natl Acad Sci U S A 111(48):17140-17145.
10. de Boer, C. G., E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev. 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol 38(1):56-65.

Figure Legends

Figure 1. *Protein-DNA binding is not limited to sequence-specific (A) and non-specific (B), but can be also characterised by an intermediate regime of non-consensus binding which is non-specific for random DNA sequences (C) but is characterised by increased binding strength for DNA regions enriched with certain repeated nucleotide patterns (D). The DNA lattice units shown in different colours may correspond to individual base pairs or larger regions. Different geometric shapes of such units used on the figure do not imply that such shapes are actually visible in the DNA structure – these may also correspond to alterations of the DNA double helix stability or other slight perturbations of the energy landscape.*