OUTLOOK

Designer genes courtesy of artificial intelligence

Alexander Hoffmann

Signaling Systems Laboratory, Department of Microbiology, Immunology, and Molecular Genetics, Institute for Quantitative and Computational Biosciences, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California 90025, USA

The core promoter determines not only where gene transcription initiates but also the transcriptional activity in both basal and enhancer-induced conditions. Multiple short sequence elements within the core promoter have been identified in different species, but how they function together and to what extent they are truly species-specific has remained unclear. In this issue of *Genes & Development*, Vo ngoc and colleagues (pp. 377–382) report undertaking massively parallel measurements of synthetic core promoters to generate a large data set of their activities that informs a statistical learning model to identify the sequence differences of human and *Drosophila* core promoters. This machine learning model was then applied to design gene core promoters that are particularly specific for the human transcriptional machinery.

The core promoter is the determinant for accurate initiation (at the +1 transcription start site [TSS]) by RNA polymerase II. Its most prominent sequence element is the TATA box at −25, which is bound by the TATA binding protein (TBP), a key DNA binding subunit of the TFIID complex (Hoffmann et al. 1990). However, as early as the 1980s, both mutagenesis and DNase1 footprinting studies indicated functional interactions by transcription initiation factors with a much larger DNA segment, extending all the way to +35 (Van Dyke et al. 1988). What was intriguing was the realization that such interactions with the downstream region correlated with the transactivation functions of transcriptional activators. Because of an apparent lack of consensus sequence in that downstream region and histone homologies in TFIID subunits (Hoffmann et al. 1996), TFIID was described as a specialized nucleosome that commits the transcription initiation site (Hoffmann et al. 1997), but that turned out to be far-fetched.

Despite the functional evidence of physical interactions by TFIID subunits in the downstream promoter region, the sequence elements directing these interactions remained elusive. In mammalian promoters, a highly heterogeneous initiator (INR) sequence was identified (Smale and Baltimore 1989), and in *Drosophila*, the Kadonaga group (Burke and Kadonaga 1996) showed that a downstream promoter element (DPE) is important for promoter activity. Combined with the more recently identified "motif ten element" (Lim et al. 2004), this was renamed the downstream promoter region (DPR). However, whether these human and *Drosophila* promoter elements are truly species-specific or have functional analogs and whether they function independently or interdependently to direct transcriptional initiation remained enigmatic for many years, while the research field turned to elucidating enhancer function and chromatin architecture.

The Kadonaga laboratory (Vo ngoc et al. 2023) recently took a fresh approach to determining the functionality of sequences within the core promoter by developing a massively parallel reporter assay (MPRA) that leverages their expertise in in vitro transcription with RNA analysis by next-generation sequencing. Their HARPE method (high-throughput analysis of randomized promoter elements), applied to human transcription-competent nuclear extract, generated data on ~200,000 promoter variants. This is a tiny fraction of the potential DPR sequence space of 4^{30} possible variants but is sufficient to train a machine learning model (Vo Ngoc et al. 2020). With the help of such statistical learning algorithms, which are excellent at identifying patterns associated with a specified condition (e.g., high transcriptional activity) and are often referred to as artificial intelligence, they determined what an effective DPR was in human cells, especially for genes lacking a TATA box. This was an important finding because it provided evidence for the functionality of the physical interactions observed in earlier studies with human extracts.

[*Keywords*: transcription; RNA polymerase II; core promoter; gene expression; *Drosophila*]
Corresponding author: ahoffmann@ucla.edu
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.350783.123. Freely available online through the *Genes & Development* Open Access option.

In the present study, Vo ngoc et al. (2023) asked whether and how the human and *Drosophila* DPRs were different. To this end, they used the HARPE method not in human but in *Drosophila* transcription-competent extracts, generating data on ~200,000 promoter variants. They showed that sequence variation in the DPR was a key determinant of promoter activity and used the data to train a machine learning model to identify sequences to determine *Drosophila* DPR activity. Comparing the sequence patterns identified by the human and *Drosophila* machine learning models, they identified a similar sequence motif of RGWYS that in human-specific variants is often shifted 1 nt upstream to +27 to +31, while in *Drosophila* it is optimally at +28 to +32. They then experimentally tested and confirmed that the 1-nt shift was the cause of species-specific activity of the same promoter construct.

However, other sequence features beyond the +1 shift may also reflect species differences accumulated over 800 million years of evolution that separate flies and humans. Machine learning algorithms are not only well suited to identify patterns that distinguish two conditions but may also be used to generate patterns. The recently released ChatGPT is an example of a machine learning model that generates new speech based on the patterns that it has identified in a vast training set of human speech. In this application, the Kadonaga laboratory (Vo ngoc et al. 2023) applied this ability to predict sequences that are maximally human- or *Drosophila*-specific. They then tested predicted variants experimentally in human or *Drosophila* extracts and confirmed that while both humans and *Drosophila* have similar DPR sequences, they are subtly distinct. Their work demonstrated that such subtle distinctions are identifiable using artificial intelligence models, which in turn can be used to design genes with specified gene regulatory design criteria.

The present study charts out an exciting new approach that could develop into a highly productive toolbox for addressing long-standing questions about transcriptional control mechanisms that have not been answered by even the highest-resolution molecular biophysical approaches: How does the core promoter determine gene expression inducibility, the gene's responsiveness to specific transcription factors? While chromatin architecture and looping are clearly important, there are also sequence determinants within the core promoter, but we do not yet understand what they are. Indeed, those core promoter sequence determinants may be different for different classes of transcription factors (classified by their activation domains, coactivators, or activation mechanisms) or their locations. Furthermore, the machine learning models for core promoters will surely find utility in the design of gene constructs for the expanding range of gene- and cell-based therapeutic approaches. A key question may be how well predictions based on in vitro naked DNA templates translate to gene expression control in the native chromatin context or whether the HARPE method or another MPRA may be designed for nucleosomal DNA or native chromatin contexts.

Ultimately, of course we are not satisfied with the amazing ability to predict designer genes with artificial intelligence, but we are aiming for a biophysical understanding. Indeed, there has been dramatic progress in the structures and the molecular interactions of TFIID (Patel et al. 2018) and the preinitiation complex (Chen and Xu 2022), but future studies may also aim to account for the dynamics and the bursts of initiation, such as their frequency and size (Larsson et al. 2019). The sequence preferences identified by Vo ngoc et al. (2023) may be a useful tool in addressing such questions and may further fuel the renaissance of biophysical studies of transcriptional control.

Acknowledgments

I thank Michael Carey for always insightful discussions about gene expression control, and acknowledge recent research funding support from National Institutes of Health grants R01AI127864, R01AI132731, R01AI127867, and R01AI132835.

References

Burke TW, Kadonaga JT. 1996. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **10:** 711–724. doi:10.1101/gad.10.6.711

Chen X, Xu Y. 2022. Structural insights into assembly of transcription preinitiation complex. *Curr Opin Struct Biol* **75:** 102404. doi:10.1016/j.sbi.2022.102404

Hoffmann A, Sinn E, Yamamoto T, Wang J, Roy A, Horikoshi M, Roeder RG. 1990. Highly conserved core domain and unique N terminus with presumptive regulatory motifs in a human TATA factor (TFIID). *Nature* **346:** 387–390. doi:10.1038/346387a0

Hoffmann A, Chiang CM, Oelgeschläger T, Xie X, Burley SK, Nakatani Y, Roeder RG. 1996. A histone octamer-like structure within TFIID. *Nature* **380:** 356–359. doi:10.1038/380356a0

Hoffmann A, Oelgeschläger T, Roeder RG. 1997. Considerations of transcriptional control mechanisms: do TFIID–core promoter complexes recapitulate nucleosome-like functions? *Proc Natl Acad Sci* **94:** 8928–8935. doi:10.1073/pnas.94.17.8928

Larsson AJM, Johnsson P, Hagemann-Jensen M, Hartmanis L, Faridani OR, Reinius B, Segerstolpe Å, Rivera CM, Ren B, Sandberg R. 2019. Genomic encoding of transcriptional burst kinetics. *Nature* **565:** 251–254. doi:10.1038/s41586-018-0836-1

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. 2004. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **18:** 1606–1617. doi:10.1101/gad.1193404

Patel AB, Louder RK, Greber BJ, Grünberg S, Luo J, Fang J, Liu Y, Ranish J, Hahn S, Nogales E. 2018. Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science* **362:** eaau8872. doi:10.1126/science.aau8872

Smale ST, Baltimore D. 1989. The 'initiator' as a transcription control element. *Cell* **57:** 103–113. doi:10.1016/0092-8674(89)90176-1

Van Dyke MW, Roeder RG, Sawadogo M. 1988. Physical analysis of transcription preinitiation complex assembly on a class II gene promoter. *Science* **241:** 1335–1338. doi:10.1126/science.3413495

Vo Ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT. 2020. Identification of the human DPR core promoter element using machine learning. *Nature* **585:** 459–463. doi:10.1038/s41586-020-2689-7

Vo Ngoc L, Rhyne TE, Kadonaga JT. 2023. Analysis of the *Drosophila* and human DPR elements reveals a distinct human variant whose specificity can be enhanced by machine learning. *Genes Dev* (this issue). doi:10.1101/gad.350572.123