Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence

Justin B. Kinney[123], Anand Murugan[3], Curtis G. Callan, Jr.[3], and Edward C. Cox[4]

A biophysical understanding of how protein-DNA and protein-protein interactions allow cells to regulate transcription has been limited by the lack of methods for measuring these interactions at specific promoters and enhancers *in vivo*. Here we show how a simple experiment, in which a library of partially mutated regulatory sequences are partitioned according to their transcriptional activities and then sequenced *en masse*, can reveal quantitative information about the biophysical interactions that allow a specific regulatory sequence to function. Our approach [1] provides a generally applicable method for characterizing biophysical mechanisms of transcriptional regulation in living cells.

***Keywords* — transcriptional regulation, deep sequencing, thermodynamic models, mutual information**

I. Background

Understanding how transcriptional regulatory sequences (TRSs) use different arrangements of protein binding sites to encode regulatory programs remains a major challenge for molecular biology. High-throughput methods have spurred great progress in cataloging the genome-wide distribution of binding sites and the sequence-specificities of individual regulatory proteins. However, determining how a specific TRS integrates information from multiple DNA-bound proteins still requires a laborious series of biochemical experiments that typically provide only qualitative information [2]. A biophysical understanding of the transcriptional regulatory code will therefore require new quantitative experimental techniques for characterizing how individual TRSs function *in vivo*.

II. Experimental design and analysis results

We hypothesized that, by measuring the activities of a large number of TRSs containing scattered point mutations, we would be able to characterize the protein-DNA and protein-protein interactions that allow a specific TRS to function *in vivo*. Our reasoning was that point substitution mutations tend to alter protein-DNA binding energies while maintaining the spatial arrangement of binding sites. By quantitatively modeling how mutation-induced changes in the DNA-binding energies of different proteins affect transcription, we would therefore be able to characterize the protein-DNA and protein-protein interactions through which a specific TRS regulates transcription.

We applied this mutagenesis-based approach to the *Escherichia coli lac* promoter. Fluorescence-activated cell sorting and 454 pyrosequencing was used to characterize the activities of ~200,000 *lac* promoters mutagenized in a 75 bp region that contains binding sites for RNA polymerase (RNAP) and the transcription factor CRP. A thermodynamic model of how the DNA sequence of this region affects transcription was then fit to the resulting sequence data. In this way, we determined the sequence-dependent binding energy of both CRP and RNAP *de novo*. We also inferred the *in vivo* interaction energy between these two proteins, achieving near agreement with a previous measurement [3].

A recently identified relationship between likelihood and mutual information [4] allowed us to do this inference without assuming a quantitative model of experimental noise. Freedom from having to independently characterize experimental noise was critically important: it enabled us to learn much more about *in vivo* biophysics from a large number of noisy measurements (obtained using deep sequencing) than would have been possible using a necessarily much smaller number of precise measurements.

III. Conclusion

Deep sequencing can be used to measure protein-DNA and protein-protein interaction energies in living cells. This ability should be useful for addressing many different questions in molecular biology.

References

[1] Kinney JB, Murugan A, Callan CG, Cox EC (in press) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci.*

[2] Carey MF, Peterson CL, Smale ST (2009) *Transcriptional regulation in eukaryotes: concepts, strategies, and techniques* (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

[3] Kuhlman T, Zhang Z, Saier MH, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA* 104:6043-6048.

[4] Kinney JB, Tkačik G, Callan CG (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* 104:501-506.

Acknowledgements: This work was supported in part by NSF grant PHY-0650617, NIH grants GM078591 and GM071508, DOE grant DE-FG02-91ER40671, and the Simons Foundation.

[1]Cold Spring Harbor Laboratory. E-mail: jkinney@cshl.edu

[2]Lewis-Sigler Institute for Integrative Genomics, Princeton University.

[3]Department of Physics, Princeton University.

[4]Department of Molecular Biology, Princeton University.