Binding-Site Affinity Modeling of Positional Dependencies and Context-Sensitive Nucleotide Insertions and Deletions

Todd R. Riley[1,3], and Harmen J. Bussemaker[2,3]

***Short Abstract* — Accurate quantification of the binding specificity of transcription factors (TFs) is crucial for understanding their biological function. Current models for predicting binding affinity from DNA sequence typically assume positional independence and fixed binding-site geometry. However, it has been shown that these assumptions are often invalid. New high-throughput technologies for profiling *in vitro* TF-DNA interaction provide an opportunity for directly inferring free-energy parameters related to positional dependencies and context-sensitive nucleotide insertions and deletions within a binding site.**

***Keywords* — Transcription factor (TF), TF-DNA interactions, sequence-specific binding affinity, weight matrix, MatrixREDUCE, PSAM, PHMM.**

I. Introduction

THE regulation of gene expression by transcription factors (TFs) is of paramount importance to the overall control of cell function. However, our current understanding of the sequence specificity of TFs is limited. Current models typically assume that each nucleotide position in a putative binding site contributes independently to the overall binding affinity of the TF for the site [1,2]. In addition, these models assume that the residue-nucleotide binding interaction geometry is static for all possible nucleotide sequences, and that consequently all the binding-sites are of equal length [1,2]. However, analysis has shown that for some TF-DNA interactions the positional-independence assumption is not valid [3,4], and that some TF-DNA interactions tolerate nucleotide insertions and deletions relative to the consensus motif [5]. Subtle differences in binding specificity between TFs can lead to qualitative differences in the downstream processes they control. [6]. It is therefore crucial to develop accurate quantitative models for predicting TF binding affinity landscapes from genome sequence.

Discussion

We have developed an extension to the biophysical model underlying the MatrixREDUCE algorithm [2] that detects deviations from the position-specific affinity matrix (PSAM) model [2] due to dinucleotide dependencies and tolerated context-sensitive nucleotide insertions and deletions. Our new model pinpoints exactly where in the binding site the positional-independence assumption breaks down. In addition, it estimates the energetic costs of context-sensitive nucleotide insertions and deletions within a half-site and within variable-length spacers between half-sites.

II. Conclusion

By applying a quantitative biophysical modeling approach to high-throughput *in vitro* binding data, we are able to build more accurate models of TF binding specificity.

References

[1] Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **23**, 109-113.

[2] Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141-e149.

[3] Roulet E, et al. (2002) High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotech.* **20**, 831-835.

[4] Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities, *Nature Biotech.* **11**, 1429-1435.

[5] Riley TR, et al. (2009) The P53HMM Algorithm: using Profile Hidden Markov Models to detect p53-responsive genes. *Bioinformatics* 2009, 10:111.

[6] Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233-237

Acknowledgments: This work was funded by NIH grant HG003008.

[1]Department of Biological Sciences, Columbia University, New York, NY 10027. E-mail: tr2261@columbia.edu

[2]Department of Biological Sciences, Columbia University, New York, NY 10027. E-mail: hjb2004@columbia.edu

[3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032