Statistical Mechanical Modeling of Protein-DNA Interactions

Allan Lazarovici[1] and Harmen Bussemaker[2]

Transcription factors (TFs) are known to play an essential role in the regulation of gene expression. We have developed a computational method to model the competitive binding of transcription factors to DNA using equilibrium statistical mechanical techniques. When fit to in vivo data (eg., ChIP-chip), the model can infer cell-state-specific TF concentrations. Our framework is quite general, and we are presently working towards adapting it to single nucleotide resolution data from next-generation sequencing experiments.

I. Purpose

TRANSCRIPTION factors help regulate gene expression by binding to either enhancer or promoter regions of DNA. In order to learn about the nucleotide sequence binding preferences of TFs, scientists have in recent years conducted both *in vitro* and *in vivo* microarray experiments. While the *in vitro* binding of purified TFs to naked DNA is now reasonably well understood [1], predicting TF occupancy from genome sequence has proven to be much more difficult, due to the presence of other TFs and histones.

We have developed a computational method that attempts to address the challenges posed by analyzing *in vivo* data. The inputs to our model are experimental data (eg. ChIP-chip) and previously known binding preferences of factors (eg TFs and nucleosomes). Our model works by evaluating the probability of observing every possible configuration of factors on the DNA sequence. These probabilities are a function of factor binding preferences and factor concentrations, the latter of which are obtained by performing a nonlinear fit to the experimental data. Our framework is quite general in the sense that the user is allowed to choose which N factors are to be included in the model, where N is greater than or equal to one.

Our methodology differs from several other published models in a couple of respects [3-4]: (i) TF concentration is inferred from experimental data (ii) TFs and nucleosomes are modeled together.

Using synthetic data, we have shown that our model can account for direct TF-TF competition as well as nucleosome-mediated cooperativity.

As a first step towards analyzing *in vivo* data, we decided to examine the role that saturation played in the Harbison ChIP-chip experiments [2]. This question piqued our curiosity because in an earlier model developed in our lab, the TF concentration is assumed to be well below saturation [3]. When fitting a model consisting of a single TF to the Harbison data, accounting for saturation leads to a significant improvement in the quality of fit. This improvement can also be observed when our model is used to analyze *in vitro* data.

In addition to computing the probability of observing every possible factor configuration, our model can answer related questions eg. the probability of observing a bound factor at a specified nucleotide position and the probability that a position of genome sequence is unbound *in vivo*.

Next-generation sequencing experiments (eg Chip-Seq), which offer single nucleotide resolution of transcription factor binding, are perhaps one of the best ways to answer the questions posed above. By adapting our framework to analyze such data, we should learn a great deal about the strengths and limitations of our approach.

References

[1] Zhu C, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research* 2009. **19**: 556-566

[2] Harbison, C, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004. **431**, 99-104.

[3] Segal, E, et al. Predicting expression patters from regulatory sequence in Drosophila segmentation. *Nature*, 2009. **451**, 535-540.

[4] Granek, J and Clarke N. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology*, 2005. **6**:R87.

[5] Foat B, Morozov A and Bussemaker H. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 2006. **22(14)** 141-149.

[1] Department of Electrical Engineering, Columbia University, USA. E-mail: allanl@ee.columbia.edu

[2] Department of Biological Sciences, Columbia University, USA. E-mail: hjb2004@columbia.edu