

From Promoter Sequence to Expression: A Probabilistic Framework -Award Winning Poster

Segal, E.¹, Barash, Y.², Simon, I.³, Friedman, N.² and Koller, D.¹ ¹ Computer Science Department, Stanford University, Palo Alto, CA, USA ² School of Computer Science & Engineering, Hebrew University of Jerusalem ³ Whitehead Institute, Cambridge, Mass, USA

We present a probabilistic framework that models the process by which transcriptional binding explains the mRNA expression of different genes. Our model unifies the two key components of this process: the prediction of gene regulation events from sequence motifs in the gene's promoter region, and the prediction of mRNA expression from combinations of gene regulation events in different settings. Our approach has several advantages. By learning promoter sequence motifs that are directly predictive of expression data, it can improve the identification of binding site patterns. It is also able to identify combinatorial regulation via interactions of different transcription factors. Finally, the general framework allows us to integrate additional data sources, including data from the recent binding localization assays. We demonstrate our approach on the cell cycle data of Spellman et al.(1998), combined with the binding localization information of Simon et al(2000). We show that the learned model predicts expression from sequence, and that it identifies coherent co-regulated groups with significant transcription factor motifs. It also provides valuable biological insight into the domain via these co-regulated "modules" and the combinatorial regulation effects that govern their behavior.