INFERRING BINDING SITE MOTIFS FROM HIGH-THROUGHPUT IN VITRO DATA

Yaron Orenstein[1], Ron Shamir[1]
[1] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

Understanding gene regulation is a key challenge in today's biology. The new technologies of protein binding microarrays (PBMs) and high-throughput SELEX (HT-SELEX) allow measurement of the binding intensities of one transcription factor (TF) to an enormous number of synthetic double-stranded DNA probes in a single experiment. The PBM technology is based on microarrays, while HT-SELEX uses deep sequencing. The ChIP-seq technique uses deep sequencing to identify bound DNA segments *in vivo*. A key computational challenge is inferring the binding site motif of the tested TF from the experimental data.

Recently, a new study (Jolma *et al.* Cell 2013) reported the results of hundreds of HT-SELEX experiments on human TFs, including many TFs covered by PBM technology. We assessed the similarities and differences between PBM and HT-SELEX technologies, and measured the performance of binding models produced by each technology in predicting *in vivo* binding. Using published HT-SELEX-derived models to predict PBM bound probes results in worse performance than PBM-derived models (average AUC 0.78 compared to 0.89). Average correlation between the top k-mers ranked by the two technologies is just over 0.5. HT-SELEX-derived models are slightly better in predicting *in* vivo binding (average AUC 0.72 compared to 0.7 on ChIP-seq data).

Our analysis currently focuses on measuring and correcting for biases. We observed GC-bias in the sequencing files, as well as systematic enrichment of specific k-mers. We will report progress towards the development of a robust computational pipeline to generate an accurate binding model from HT-SELEX data.