Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources

Harri Lähdesmäki, Alistair G. Rust & Ilya Shmulevich
Institute for Systems Biology, Seattle, WA, USA

Motivation

- Transcriptional regulation is a central control mechanism for many biological processes
- Relatively little is known about genome-wide binding of transcription factors (TF) to gene promoters
- Most of the previous binding site prediction tools are based on hypothesis testing (i.e., p-value)
- Many TF binding sequences are relatively short and degenerate, and hence they occur frequently in a genome just by chance

Our goal: to develop a method that

1. is probabilistic in nature and thus outputs a probability of binding
2. answers the question of whether the whole promoter has a binding site (but also outputs the probability of binding to each nucleotide position separately)
3. provides a principled way of combining multiple data sources, such as evolutionary conservation, regulatory potential, CpG islands, nucleosome positioning, DNase hypersensitive sites, ChIP-chip, and other prior knowledge, into a unified probabilistic framework

- We demonstrate that our method significantly improves TF binding predictions

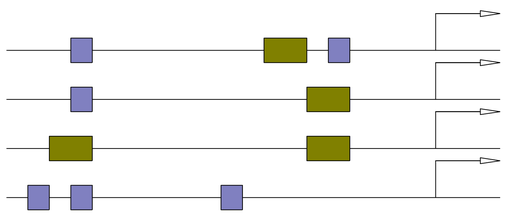


FIGURE 1: Method considers all possible binding site configurations and weights them according to their probability

Modeling Framework & Methods

- TFs are typically associated with multiple motif models (e.g. TRANSFAC position weight matrices) $\Theta = (\theta^{(1)}, \ldots, \theta^{(m)})$
- A key unknown quantity is the number of (hidden) binding sites Q in a promoter sequence S (Fig. 1)
- Given S, Θ and a background model ϕ, compute the probability of having $c = 0, 1, \ldots$ binding sites

$$P(Q = c|S, \Theta, \phi) \propto P(S|Q = c, \Theta, \phi) \times P(Q = c|\Theta, \phi)$$

- The probability of binding can be assessed as

$$P(Q > 0|S, \Theta, \phi)$$

- Assume that each additional data source is in the form of $\mathcal{D} = (P(1), \ldots, P(N))$, where $P(i)$ denotes the probability that the ith nucleotide has one of the above properties (conserved, low nucleosome occupancy, etc.)
- Given (putative) binding start locations A and configurations π, combine different data sources as

$$P(S, \mathcal{D}|A, \pi, \Theta, \phi) = P(S|A, \pi, \Theta, \phi)P(\mathcal{D}|A, \pi)$$

- The intuitive rationale for defining $P(\mathcal{D}|A, \pi)$ is to assign higher probabilities for those (A, π) that are located in regions that are more likely (in light of additional data $\mathcal{D}$) to contain functional binding sites
- Efficient recursive formula to compute

$$P(S, \mathcal{D}|Q = c, \Theta, \phi)$$

and consequently

$$P(Q = c|S, \mathcal{D}, \Theta, \phi)$$

- Similar method also in Bayesian context, but estimation requires an MCMC algorithm

Results

- To test the performance of our method, we have constructed a test set of annotated binding sites in the mouse genome from ORegAnno and ABS
- TF binding specificities are taken from TRANSFAC 10.3

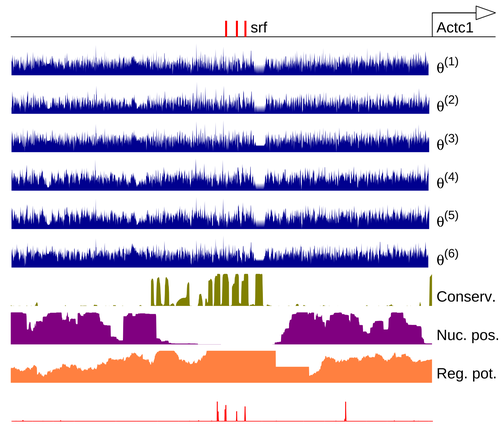
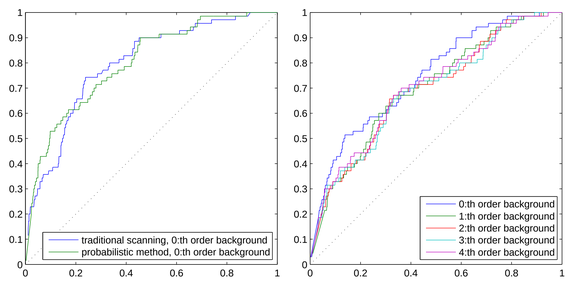


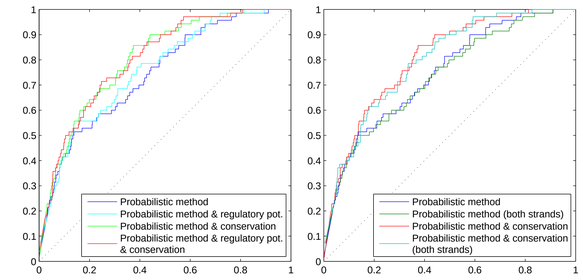
FIGURE 2: An illustrative example of data fusion. From top: annotated binding site(s), "raw matrix predictions," conservation probabilities, nucleosome occupancy probabilities, regulatory potential likelihood ratios, and estimated binding locations (i.e., probabilities)



(a) (b)

FIGURE 3: ROC curves: (a) A comparison with a traditional scanning method and (b) different Markovian background models

- Basic method (without data fusion) performs better than traditional scanning (Fig. 3 (a))
- Comparison of different Markovian background models (Fig. 3 (b))
- Combining multiple information sources clearly improves binding predictions (Fig. 4 (a))
- Method can, e.g., equally well use both strands of DNA (Fig. 4 (b))



(a) (b)

FIGURE 4: ROC curves: (a) Illustrative results for different additional data sources (b) single vs. double-stranded DNA

- Other information sources, e.g. ChIP-chip data, can be easily included

Conclusions & Future Directions

- Method provides a principled probabilistic integration of multiple data sources
- Regularization via various informative priors
- Probabilistic approach provides an intuitive interpretation
- Straightforward to integrate with other probabilistic/Bayesian methods

References

[1] Lähdesmäki, H., Rust, A. G. and Shmulevich, I. (Manuscript in preparation) Probabilistic Inference of Transcription Factor Binding from Multiple Data Sources.