Probabilistic transcription factor binding prediction: joint inference from multiple data sources

Harri Lähdesmäki[1,2], Alistair G. Rust[1] & Ilya Shmulevich[1]

[1]Institute for Systems Biology, Seattle, WA, USA
[2]Institute of Signal Processing, Tampere University of Technology, Tampere, FINLAND

Motivation

- Transcriptional regulation is a central control mechanism for many biological processes
- Many TF binding sequences are relatively short and non-unique, 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 can also output the probability of binding to each nucleotide position separately)
3. provides a principled way of combining multiple data sources at the genome level, such as multiple motif models, evolutionary conservation, regulatory potential, CpG islands, nucleosome positioning, DNase hypersensitive sites, ChIP-chip, and other prior knowledge, into a unified probabilistic framework

Modeling Framework & Methods

- TFs are typically associated with multiple motif models (e.g. TRANSFAC PSWMs) $\Theta = (\theta^{(1)}, \ldots, \theta^{(m)})$
- A key unknown quantity is the number of binding sites Q in a promoter sequence S (Fig. 1)

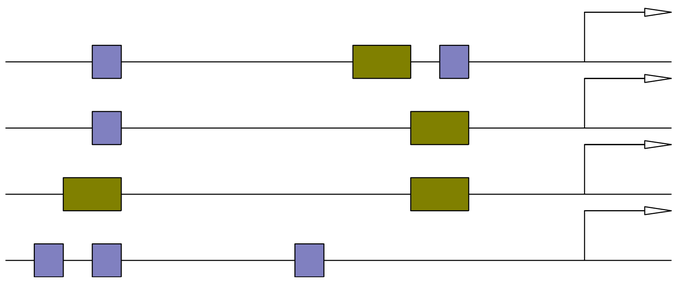

FIGURE 1: Our method considers all possible binding site configurations and weights them according to their probability

- 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
- An efficient recursive formula to compute

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

and consequently

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

- A similar method also in Bayesian context: Θ and ϕ are random variables and estimation uses an MCMC algorithm

Results

- We construct a test set of annotated binding sites in the mouse genome from ORegAnno and ABS, and demonstrate that our probabilistic data fusion significantly improves TF binding predictions
- TF binding specificities from TRANSFAC 10.3

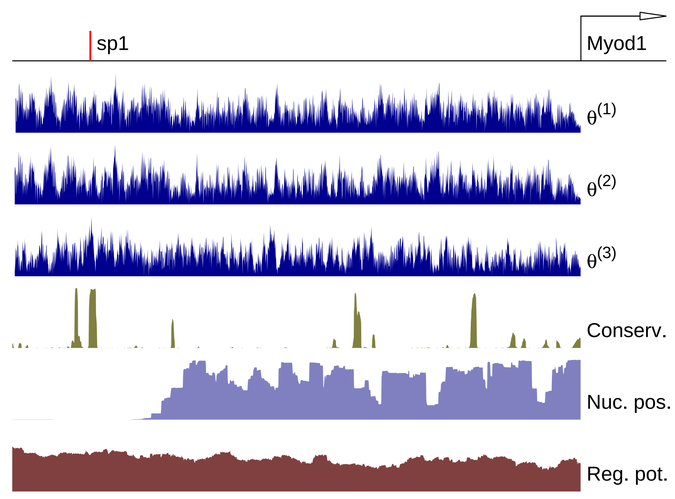

FIGURE 2: An illustrative example of data fusion. From top: annotated binding site(s), "raw matrix predictions," conservation probabilities, nucleosome occupancy probabilities, regulatory potential likelihood ratios

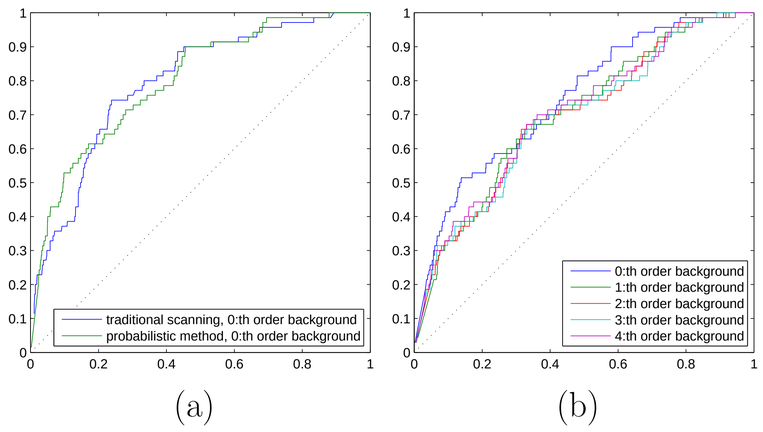

FIGURE 3: ROC curves: (a) A comparison with a traditional scanning method and (b) different Markovian background models

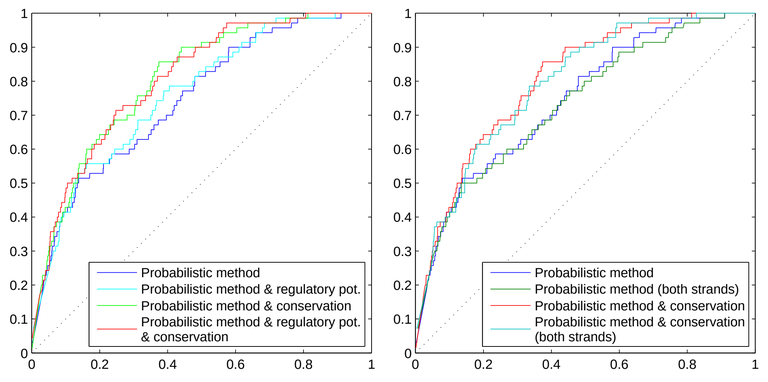

FIGURE 4: ROC curves: (a) Illustrative results for different additional data sources (b) single vs. double-stranded DNA

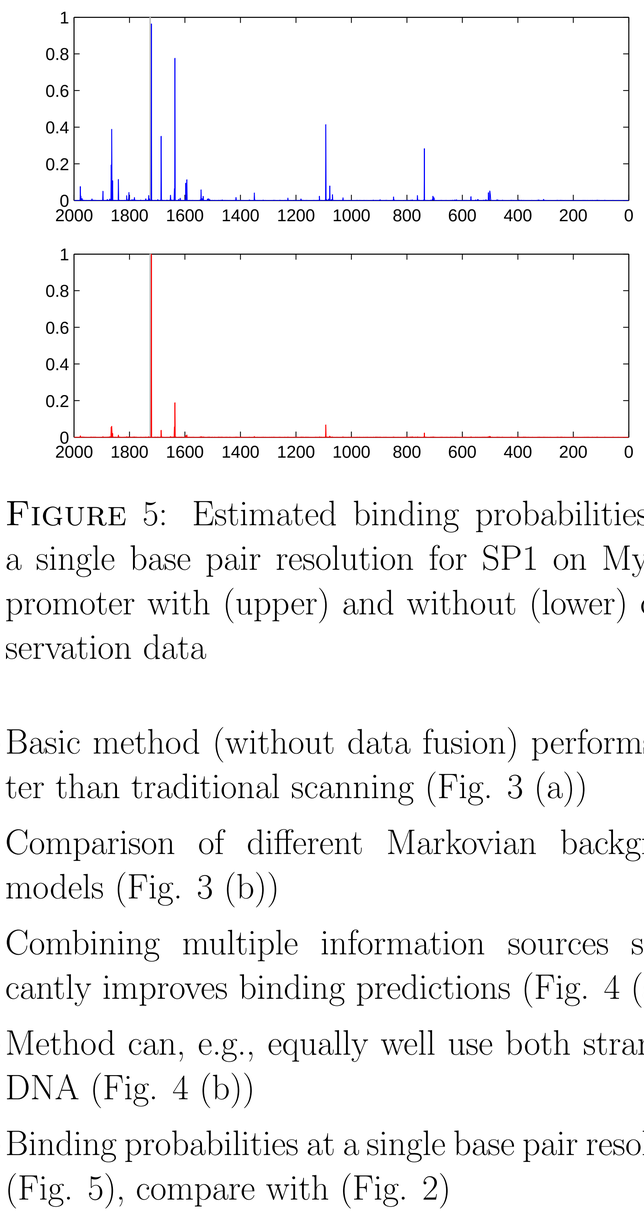

FIGURE 5: Estimated binding probabilities on a single base pair resolution for SP1 on Myod1 promoter with (upper) and without (lower) conservation data

- 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 significantly improves binding predictions (Fig. 4 (a))
- Method can, e.g., equally well use both strands of DNA (Fig. 4 (b))
- Binding probabilities at a single base pair resolution (Fig. 5), compare with (Fig. 2)

Conclusions & Future Directions

- Method provides a principled probabilistic integration of multiple data sources
- Other information sources, e.g. ChIP-chip data, can be easily included
- Straightforward to integrate with other probabilistic/Bayesian methods
- Our Bayesian method can also be used as a motif discovery method
- Can also be extended to model, e.g., combinatorial regulation

References

[1] Lähdesmäki, H., Rust, A. G. and Shmulevich, I. (Submitted) Probabilistic inference of transcription factor binding from multiple data sources.