Probabilistic data fusion for transcription factor binding prediction

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

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

Motivation

- Revealing regulatory mechanisms by which transcription factors (TF) bind and regulate gene expression is a key problem in understanding genome-wide transcriptional regulation

Our TF binding prediction method

1. provides a unified probabilistic modeling framework
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

Modeling Framework & Methods

- Multiple motif models for each TF, we use TRANSFAC (ver. 10.3) 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)

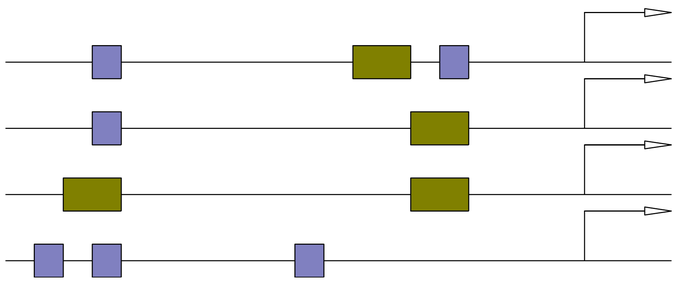


FIGURE 1: Our method considers all possible binding site configurations and weights them according to their probability

- Given S, Θ and a background model ϕ, compute the probability of having $c = 0, 1, \ldots$ binding sites

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

- Given (putative) binding locations A and configurations π, combine different data sources as

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

- $P(\mathcal{D}|A, \pi)$ assigns 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 (see also Fig. 2)
- 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 method significantly improves TF binding predictions

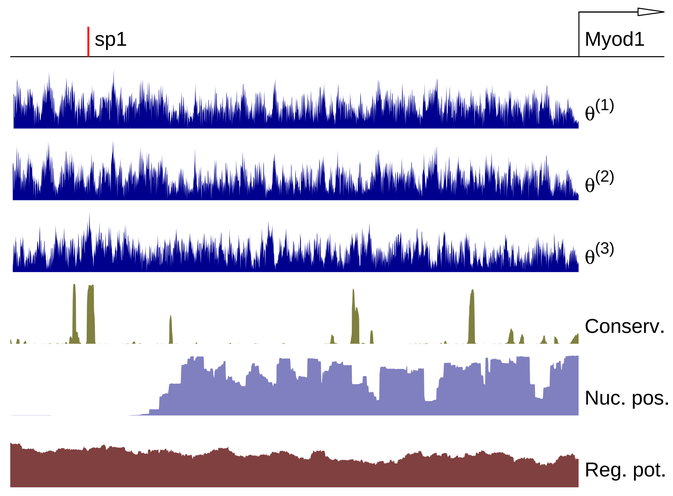
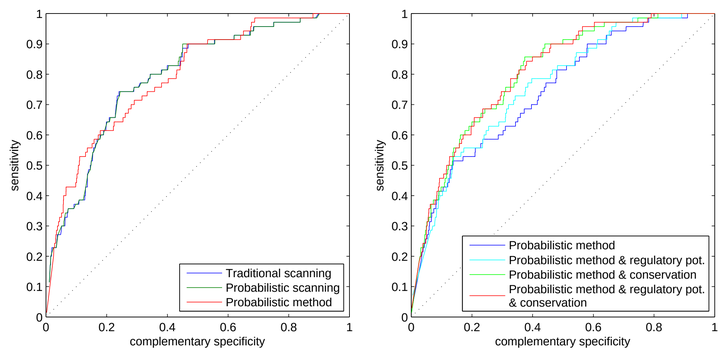


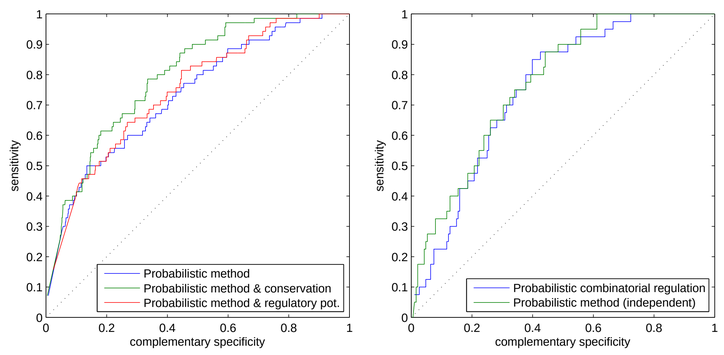
FIGURE 2: An illustrative example of data fusion. From top: annotated binding site(s), "raw matrix predictions," conservation probabilities, nucleosome occupancy probabilities, regulatory potential likelihood ratios

- Basic method (without data fusion) performs better than traditional scanning (Fig. 3 (a))
- Combining multiple information sources significantly improves binding predictions (Fig. 3 (b))
- Method can use both strands of DNA (Fig. 4 (a))
- Can easily be extended to model combinatorial regulation by multiple TFs (Fig. 4 (b))



(a) (b)

FIGURE 3: ROC curves: (a) A comparison with a traditional scanning method and (b) Illustrative results for different additional data sources



(a) (b)

FIGURE 4: ROC curves: (a) results using double-stranded DNA (b) combinatorial regulation

- Binding probabilities at a single base pair resolution (Fig. 5), compare with (Fig. 2)
- A summary of genome-wide predictions for mouse: all genes and all TFs (Fig. 6)
- ProbTF tool is publicly available (Fig. 7)

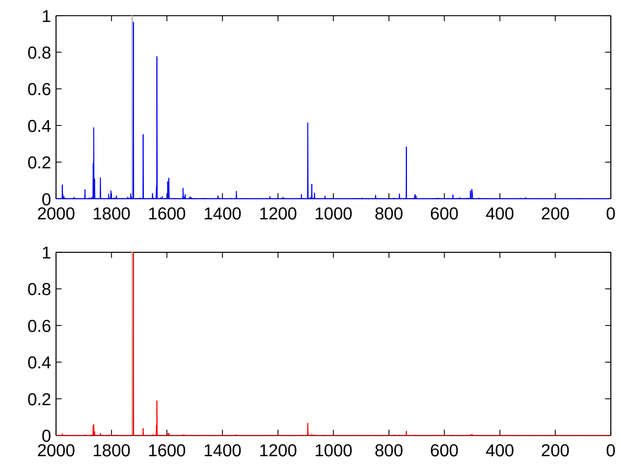
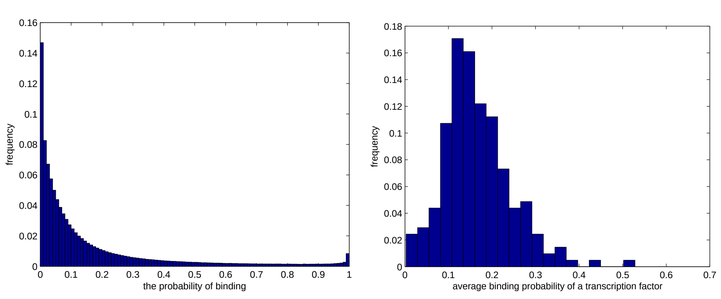


FIGURE 5: Estimated binding probabilities on a single base pair resolution for SP1 on Myod1 promoter without (upper) and with (lower) conservation data



(a) (b)

FIGURE 6: Histogram of (a) genome-wide binding predictions (b) average binding of TFs

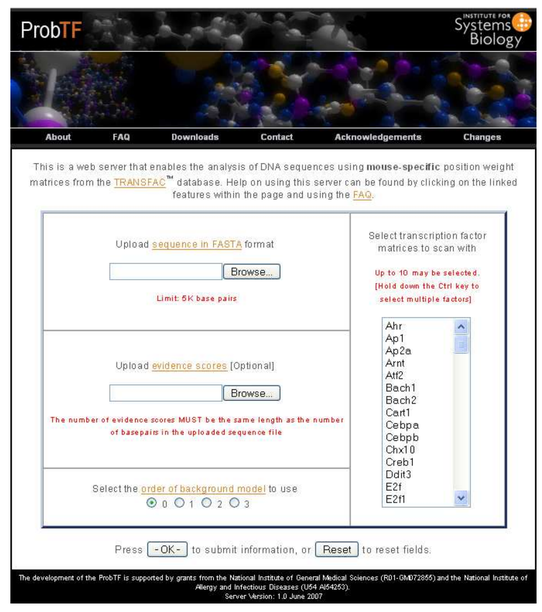


FIGURE 7: ProbTF tool available on-line at `http://www.probtf.org`

Conclusions & Future Directions

- Method provides an efficient and principled probabilistic integration of multiple data sources
- Straightforward to integrate with other probabilistic/Bayesian methods, e.g., regulatory network inference from expression data

References

[1] Lähdesmäki, H., Rust, A. G. and Shmulevich, I. (2008) Probabilistic inference of transcription factor binding from multiple data sources, *PLoS ONE*, Vol. 3, No. 3, e1820.