Inferring transcription factor targets from multiple genome-level data sources

TAMPERE UNIVERSITY OF TECHNOLOGY

Xiaofeng Dai & Harri Lähdesmäki
Department of Signal Processing, Tampere University of Technology, Tampere, Finland

Motivation

- A key problem is to reveal genome-wide regulatory mechanisms by which transcription factors (TF) bind and control gene expression

Our TF binding prediction method

1. provides a 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 practically any genome-level data: here we use multiple motif models, evolutionary conservation, regulatory potential, nucleosome positioning, and DNA duplex stability data

Modeling Framework & Methods

- Multiple binding specificity models for each TF, we use TRANSFAC PSFMs: $\Theta = (\theta^{(1)}, \ldots, \theta^{(m)})$
- A key unknown quantity is the number of binding sites Q in a promoter sequence S (Fig. 1)
- Given S, Θ and a background model ϕ, compute the probability of having $Q = 0, 1, \ldots$ binding sites

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

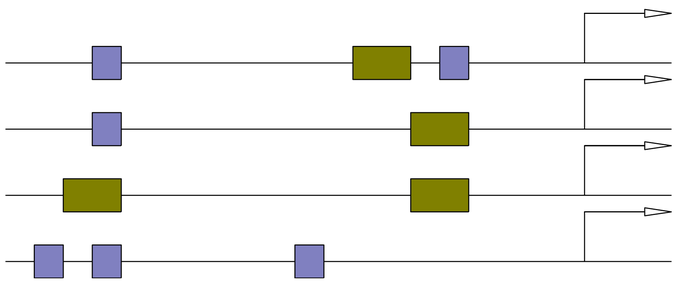


FIGURE 1: Our method considers all possible binding site configurations and weights them according to their probability

- Given (putative) binding locations A and configurations π, combine additional data $\mathcal{D}$ 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 locations (A, π) that are more likely (in light of $\mathcal{D}$) to contain functional binding sites (Fig. 2)
- Use a previously proposed method to map raw data R of a genomic location $\mathcal{X}$ into the probability of being a binding site B

$$P(\mathcal{X} \in B|R(\mathcal{X})) = \frac{P(R(\mathcal{X})|\mathcal{X} \in B)P(\mathcal{X} \in B)}{P(R(\mathcal{X}))}$$

- Combine individual data (conservation, regulatory potential, nucleosome positioning, DNA stability) into $\mathcal{D}$ using a data driven approach (Fig. 3)
- Also a Bayesian version: Θ and ϕ are random variables and estimation uses an MCMC algorithm

Results

- We use a test set of annotated binding sites in the mouse genome from ORegAnno and ABS [1] to demonstrate that our method significantly improves TF binding predictions (Fig. 2)
- Conservation and regulatory potential improve binding site prediction [1], and so do nucleosome and DNA stability data [2] (Figs. 4 & 5)

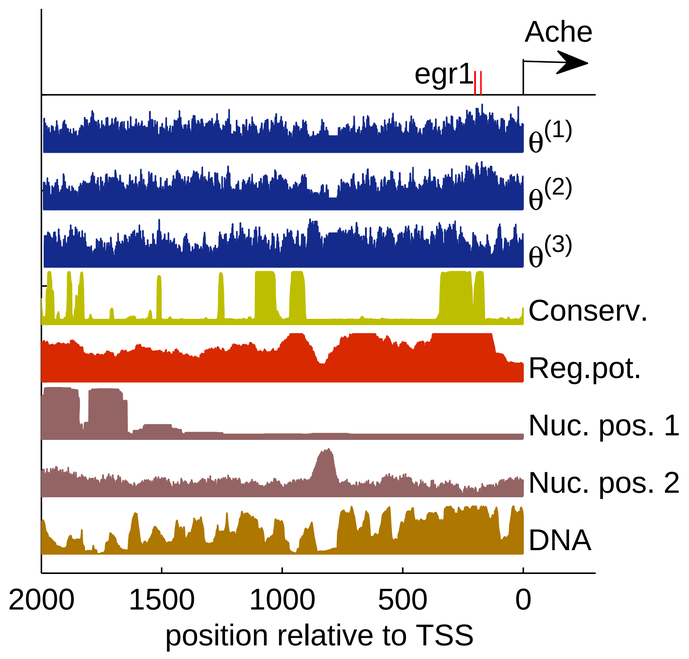


FIGURE 2: Data fusion illustration. From top: known binding site(s), PSFM predictions, conservation, regulatory potential, nucleosome positioning, DNA duplex stability

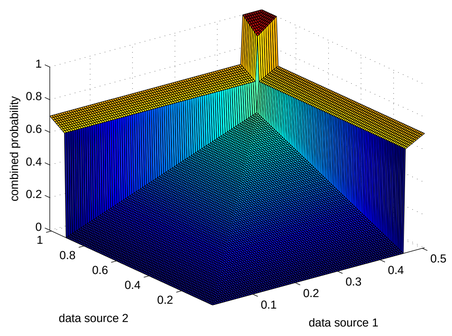


FIGURE 3: An illustration of the data driven method to combine individual data into $\mathcal{D}$.

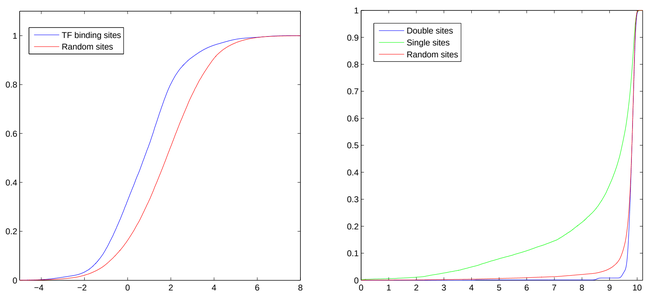


FIGURE 4: CDFs of (left) nucleosome and (right) DNA stability scores at TFBSs and random sites.

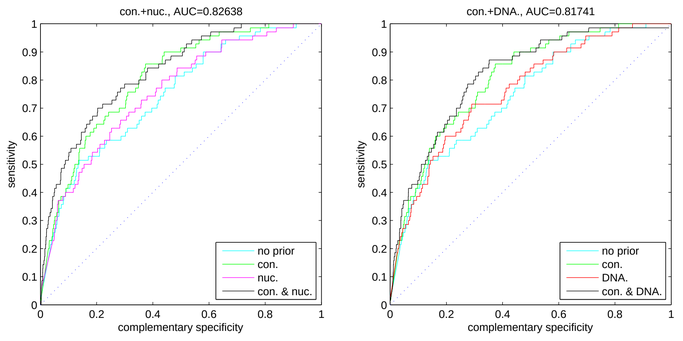
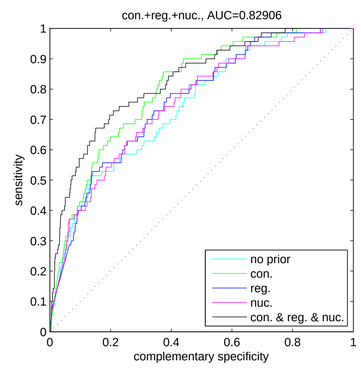


FIGURE 5: ROCs: (left) nucleosome and conservation, (right) DNA stability and conservation.

- Best results by combining binding site predictions with conservation, regulatory potential and nucleosome positioning (Fig. 6)
- Can also compute binding probabilities at a single base pair resolution (Fig. 7)
- Source code and ProbTF web tool are publicly available (Fig. 8)



Priors	AUC
no prior	0.7449
nuc.+DNA.	0.74992
DNA.(crossvalid)	0.753
nuc.	0.7555
DNA.	0.75858
reg.	0.7611
reg.+DNA.	0.77285
reg.+nuc.+DNA.	0.77413
reg.+nuc.	0.79456
con.+reg.	0.80127
con.	0.8038
con.+reg.+DNA.	0.81434
con.+reg.+nuc.+DNA.	0.81721
con.+DNA.	0.81741
con.+nuc.+DNA.	0.82526
con.+nuc.	0.82638
con.+reg.+nuc.	0.82953

FIGURE 6: (left) ROCs for combining conservation, regulatory potential and nucleosome data. (right) AUCs for all combinations.

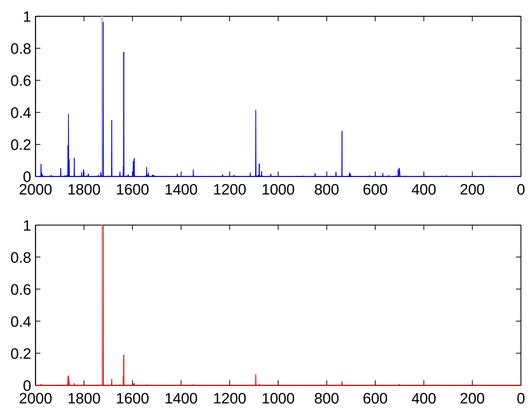


FIGURE 7: Binding probabilities at a single base pair resolution for SP1 on Myod1 promoter without (upper) and with (lower) data fusion.

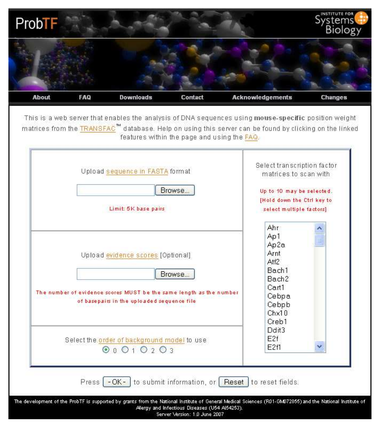


FIGURE 8: `http://www.probtf.org`

Conclusions & Future Directions

- Method provides an efficient and principled probabilistic integration of multiple data sources for TF binding site prediction
- Straightforward to integrate e.g. with regulatory network inference from expression data
- Can be easily extended to incorporate other genome-level data sources and to model combinatorial regulation by multiple TFs

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.

[2] Dai, X. and Lähdesmäki. H. Inferring transcription factor targets from multiple genome-level data sources, submitted for publication.