A probabilistic model for competitive binding of transcription factors

TAMPERE UNIVERSITY OF TECHNOLOGY

Kirsti Laurila[1] & Harri Lähdesmäki[1,2]

kirsti.laurila@tut.fi, harri.lahdesmaki@tut.fi

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Department of Information and Computer Science, Helsinki University of Technology, Finland

Motivation

Transcription factor binding site (TFBS) identification

- Experimental high-throughput techniques are laborious and one can study only certain conditions and a single transcription factor (TF) at a time
- Prediction methods needed
- Traditional methods predict binding of a single TF
  - Not realistic biologically → several TFs present in the cell, several TFBSs for different TFs on one promoter (Figure 1)
  - Good sensitivity, poor specificity
- Data fusion can improve predictions
  - Probabilistic model for TFBS prediction combining heterogenous data sources [1]
  - Good sensitivity, poor specificity
- Predictors for multiple TFs
  - Proximal binding sites
  - Cis-regulatory modules

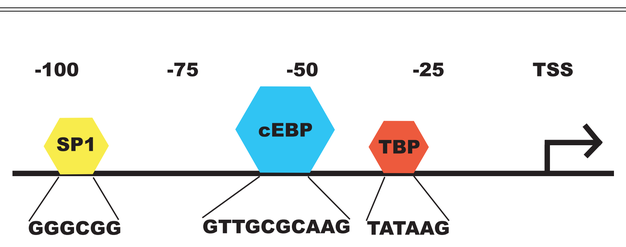

Figure 1: Known binding sites for mouse leptin (U36238) promoter. TSS=transcription starting site.

Methods

Principle behind the model

- Proposed method predicts TFBSs for multiple TFs simultaneously
- Competition between TFs
  → mimics situation in cells

Model building

- Binding is measured with probabilities
  → weak and strong binding sites
  - Background with Markovian model
  - Binding with position frequency scoring matrices
  - Dirichlet priors are used
- Bayesian posterior $P(A, \pi|S) \propto P(S|A, \pi)P(A, \pi)$
  - Set of TFs bind to specific locations on promoter
  - S=sequence
  - A= starting locations of binding sites
  - π= labels of binding TFs
- Markov Chain Monte Carlo estimation
  - Metropolis Hastings algorithm
    → proposes a new binding site or removal of an existing one

Data

- 29 mouse promoter sequences with TFBSs from ABS [2] and ORegAnno [3]
- 27 TFs, PSFMs from TRANSFAC

Results

- Comparison with compared individual predictions
- Example of mouse leptin promoter predictions in Figure 2
  - With combined individual predictions lots of false positives and overlapping predictions
    → overcome when competitive binding model is used
- ROC (receiver operating characteristic) curves in Figure 3 and AUC (area under curve) scores in Table 1

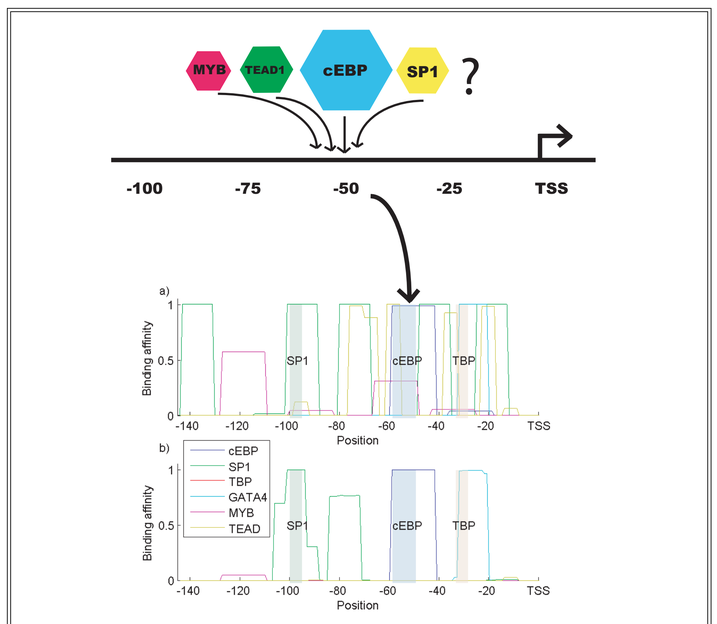

Figure 2: Predictions for mouse leptin promoter. Predictions are presented for those TFs that bind with affinity more than 0.1. Known binding sites are shaded. a) Individual predictions b) Competitive model.

Table 1: AUCs. FPR= False positive rate, Ind= Combined individual predictions, Ind norm= Combined individual predictions normalized, Comp= Competitive model.

FPR	0.10	0.20	0.30	0.40	0.50
Ind	0.0127	0.0421	0.0919	0.1524	0.2228
Ind norm	0.0104	0.0412	0.0904	0.1503	0.2213
Comp	0.0185	0.0606	0.1179	0.1849	0.2578

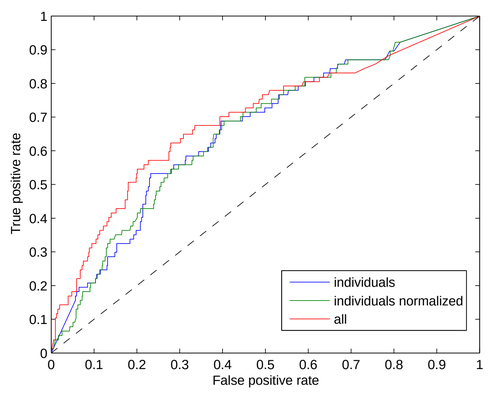

Figure 3: ROC curves for different methods.

Conclusions & Future Directions

- Number of false positives is remarkably reduced
- Non-overlapping predictions
- Existing knowledge can be used as a prior
- With Bayesian inference can some previously unpredictable binding sites be identified
- In the future incorporating protein-protein interactions to the model (Figure 4)
- Usage of also other data sources to improve predictions
  - TF concentrations
  - nucleosome positions
  - conservation data
  - ...

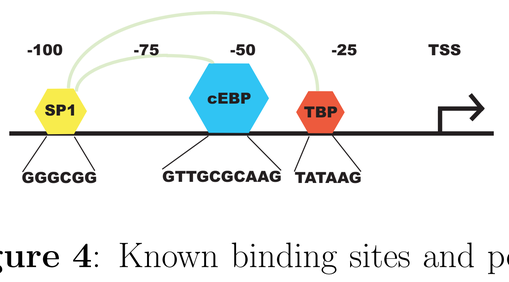

Figure 4: Known binding sites and possible interactions between TFs of mouse leptin promoter

References

[1] Lähdesmäki et al. PloS ONE, 2008.
[1] Blanco et al. Nucleic Acids Res, 2006.
[1] Griffith et al. Nucleic Acids Res, 2008.