Predicting Binding of Transcriptional Regulators with a Two-Way Latent Grouping Model

S. Kaski[1,2]**, E. Savia**[2]**, K. Puolamäki**[2]

[1]Department of Computer Science, University of Helsinki
[2]Laboratory of Computer and Information Science, Helsinki University of Technology

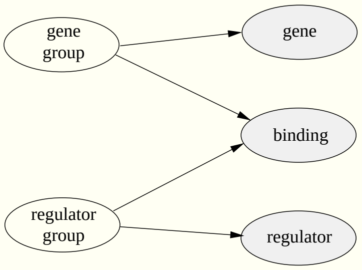
Abstract We model the binding patterns of transcription factors to the promoter regions of genes using a two-way latent grouping model. The model assumes latent gene groups and latent regulator groups and makes Bayesian prediction for the binding.

Introduction

- Binding of transcription factors to the promoter regions of genes can be measured genome-wide to reveal regulatory networks
- Measurements are expensive
- Prediction of bindings from earlier data would reduce the cost

Two-Way Latent Grouping Model

- Generative probabilistic model [Savia *et al.*, 2005]
- Assumption: latent group structure
  - Genes belong to groups of similarly behaving genes
  - Transcriptional regulators in different conditions belong to groups of similarly behaving regulators
- Probability of binding assumed to depend solely on the pair of latent gene group and latent regulator group



Data

- From genome-wide analysis of yeast [Harbison *et al.*, 2004]
- 203 DNA-binding transcriptional regulators in different conditions
- 352 location studies for 6227 genes
- Original data: P-values of the confidence of binding
- We extracted
  - high-confidence interactions (5% with the lowest P-value)
  - set where binding is the most uncertain (5% with the highest P-value)
- The rest were interpreted as missing data

Experiments

- Prediction of binding for new regulators: only 3 samples in training set
- Comparison against a state-of-the-art latent topic model *URP* [Marlin, 2004]
  - Latent group structure for the genes
  - Models each transcriptional regulator independently
- Evaluation by Gibbs sampling
- Number of groups determined using a validation set
- Baseline model: each regulator has a fixed tendency to bind, irrespective of the genes

Results

- Methods produce probabilities of binding as predictions
- Generalization into groups of genes and regulators proved to be profitable
- Number of gene groups = 2
- Number of regulator groups = 2
- Differences statistically significant (Wilcoxon signed rank test $P < 0.001$)

Method	Neg. log-likelihood	Absolute error
Two-way	0.57	0.28
URP	0.59	0.32
Baseline	1.68	0.41

Conclusion

- Feasibility study of predicting the binding patterns of transcription factors to the promoter regions of genes
- Prediction of bindings for new regulators based on earlier data works
- At best, genome-wide studies could be targeted based on a few test samples
- Two-way grouping improved prediction accuracy
- Possible extensions:
  - Including evidence from phylogenetic studies
  - Enhanced pre-processing the binding data

References

[Harbison *et al.*, 2004] Harbison, C., et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature,* **431** (7004), 99–104.

[Marlin, 2004] Marlin, B. (2004) Modeling user rating profiles for collaborative filtering. NIPS 16.

[Savia *et al.*, 2005] Savia, E., et al. (2005). Two-way latent grouping model for user preference prediction. UAI'05.