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

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

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

INTRODUCTION

Binding of transcription factors to the promoter regions of genes can be measured genome-wide to reveal regulatory networks. The measurements are expensive, however, and methods for predicting bindings from earlier data would reduce the cost. At best, genome-wide studies could be targeted based on a few test samples.

We model the binding patterns using recent ideas from collaborative filtering and biclustering. A main difference from biclustering is that we compute a Bayesian prediction using all possible clusterings.

DATA

The data comes from genome-wide location analysis of 203 DNA-binding transcriptional regulators (TR) of yeast in different rich media conditions and other environmental conditions (Harbison *et al.*, 2004); alltogether 352 location studies. The number of genes was 6227. In the original data the interactions between regulators and promoter regions were expressed as P-values of the confidence of binding. We used the P-values to extract a set of high-confidence promoter-TR interactions (5% of the data with the lowest P-value) and a set where binding is the most uncertain (5% with the highest P-value). The rest are interpreted as missing data (not known whether they bind); this is a conservative choice that will be optimized later. The result is a binary matrix with 90% missing values.

METHODS

We assumed the data to have a latent group structure where both the genes and the transcriptional regulators in different conditions belong to groups of similarly behaving genes or regulators. Each pair of groups has a typical binding tendency; the generative probabilistic model (Puolamäki *et al.*, 2004) assumes that the probability of binding depends solely on the latent gene group and the latent regulator group.

The results were compared against a state-of-the-art latent topic model *URP* (Marlin, 2004). It assumes the same kind of latent group structure for the genes but models each transcriptional regulator independently. Predictions were computed for both models by Gibbs sampling, with a fixed number of groups determined using a validation set. Baseline for the comparisons came from a model where each regulator has a fixed tendency to bind, irrespective of the genes, and the tendency is estimated as the frequency of binding in the training set.

EXPERIMENTS AND RESULTS

Since we expect main applications in predicting bindings for potential new regulators (or genes) where only few or very noisy measurements exist so far, we designed the experimental setting accordingly. The learning set consisted of all data, except that for 50 of the regulators only 3 bindings were included. These represent the new regulators. The rest of the bindings for the new regulators formed the test set (roughly 29,000 samples in total).

The numbers of latent groups were selected according to the results of a validation set of another 50 regulators. The optimal number of gene groups was 2 for URP and both numbers of groups were 2 for our method. Note that the numbers are small since generalization from only 3 known samples requires heavy oversimplification.

Both methods produce probabilities of binding as predictions. The average negative log-likelihood was 0.57 for our model, 0.59 for URP, and 1.68 for the baseline method. The average absolute error was 0.28 for our model, 0.32 for URP, and 0.41 for the baseline method. The differences in both measures between all the methods were statistically significant (Paired T-test $P < 0.001$).

DISCUSSION

This is a feasibility study with the obvious extensions of including other evidence about binding, for instance from phylogenetic studies, and pre-processing the binding data more carefully into a binary matrix, or treating it as real-valued data.

REFERENCES

Harbison, C., et al. (2004) *Nature,* **431** (7004), 99–104.

Marlin, B. (2004) NIPS 16.

Puolamäki, K., et al. (2004). Tech Rep A80, Publications in Computer and Information Science, Helsinki Univ Tech