Modelling Transcriptional Regulation with a Variational Bayesian Mixture of Factor Analyzers

Kuang Lin and **Dirk Husmeier**
Biomathematics and Statistics Scotland (BioSS)
Edinburgh, United Kingdom

December 4, 2007

INTRODUCTION

Quantitative modelling of the regulatory networks of the cell is one of the central challenges in computational biology. One of the most important cellular regulation mechanisms is at the transcriptional level, where the expression of a gene is controlled by the binding of diverse regulatory proteins called transcription factors (TFs) to specific DNA sequences in the promoter region of the gene. The objective of the present study is the quantitative modelling of these processes. Our approach aims to integrate gene expression profiles with transcription factor binding data, such as binding motifs in promoter regions or p-values from immunoprecipitation experiments. The model we employ is a mixture of factor analysers [1], in which the latent variables correspond to different transcription factors, grouped into complexes or modules. We pursue inference in a Bayesian framework, using the Variational Bayesian Expectation Maximization (VBEM) algorithm for approximate inference of the posterior distributions of the model parameters, and estimation of a lower bound on the marginal likelihood for model selection.

METHOD

We have investigated the application of Bayesian mixtures of factor analyzers (MFA-VBEM) to modelling transcriptional regulation in cells. Like recent approaches based on Bayesian factor analysis applied to the same problem [4, 5], MFA-VBEM allows for the fact that TFs are often subject to post-translational modifications and that their true activities are therefore usually unknown. A shortcoming of Bayesian factor analysis is the fact that it ignores interactions between TFs. This limitation is addressed by our approach: different from Bayesian factor analysis, the mixture of factor analyzers approach allows for the cooperation between TFs to form cis-regulatory modules, which is particularly common in higher eukaryotes. Our approach systematically integrates gene expression data with TF binding data. As opposed to the partial least squares (PLS) approach of [2], MFA-VBEM is a probabilistic model that allows for the noise inherent in the TF binding data. This addresses a major shortcoming of the PLS approach, where the inability to deal with measurement errors has been found to adversely affect the activity profile reconstruction accuracy. A further advantage over Bayesian factor analysis is the fact that gene expression and TF binding data are treated on an equal footing; this avoids the rather artificial division of treating gene expression profiles as data, and TF binding profiles as prior knowledge.

EVALUATION

We have evaluated the performance of the proposed method on three criteria: activity profile reconstruction, gene clustering, and network inference. The objective of the first criterion is to assess whether the activity profiles of the transcriptional regulatory modules can be reconstructed from gene expression data. The second criterion tests whether the method can discover biologically meaningful groupings of genes, indicated by significant enrichment for known gene ontologies. The third criterion addresses the question of whether the proposed scheme can make a useful contribution to computational systems biology, where one is interested in the reconstruction of gene regulatory networks from diverse sources of postgenomic data.

RESULTS

Using a synthetic data set, we found that MFA-VBEM reconstructed the hidden activity profiles of the cis-regulatory modules more accurately than PLS [2] and Bayesian factor analysis with Gibbs sampling [4]. Using gene expression profiles and TF binding profiles for *S. cerevisiae*, MFA-VBEM found biologically more plausible gene clusters than K-means, hierarchical agglomerative average linkage clustering and COSA [3], as indicated by the increased enrichment for known gene ontology terms. For the regulatory network reconstruction task, MFA-VBEM outperformed Bayesian and non-Bayesian factor analysis models on gene expression and TF binding profiles from both *S. cerevisiae* and a synthetic simulation.

Acknowledgments

This work was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

References

[1] Matthew J Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[2] Anne-Laure Boulesteix and Korbinian Strimmer. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor Biol Med Model*, 2(1):23, 2005.

[3] Jerome H Friedman and Jacqueline J Meulman. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B*, 66(4):815–49, 2004.

[4] Chiara Sabatti and Gareth M James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–46, 2006.

[5] Guido Sanguinetti, Magnus Rattray, and Neil D Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22(14):1753–9, 2006.