Poster L-12

Probabilistic Inference of transcription factor concentrations and gene-specific regulatory activities

Authors:
Guido Sanguinetti (*University of Sheffield*)
Neil Lawrence (*University of Sheffield*)
Magnus Rattray (*University of Manchester*)

Short Abstract: We present a probabilistic dynamical model to infer active transcription factor concentrations and their regulatory effect for regulatory networks of known architecture. The advantage of the model is that it allows to estimate separately the strength and sign of the regulator interactions. The model is validated on two yeast datasets.

Long Abstract:
Quantitative modelling of the regulatory network of the cell is one of the grand challenges of bioinformatics. Although recent techniques such as Chromatine Immunoprecipitation (ChIP) have uncovered much information about the architecture of the networks, any quantitative model would require the knowledge of both the concentration of transcription factor proteins at a given time and the intensity with which they can promote or repress transcription of their target genes. Experimental estimation of these variables meets with insurmountable obstacles: measuring protein concentrations is a notoriously difficult task, and little help can be gleaned from knowledge of transcription factor gene expression levels. Transcription factors are often post-transcriptionally regulated and have low and noisy expression levels. Furthermore, the effect a transcription factor has on a target gene depends greatly on the experimental conditions, making experimental estimation of the strength of regulatory relationships a difficult task.

An idea that has gained a lot of interest in recent years has been to infer information about regulatory activity from the expression levels of target genes. Using information from ChIP experiments about the structure of the network and genome-wide microarray data for the expression levels of the targets, it should be possible to gain insights on the activity of the transcription factors. Most methods aim to infer a matrix of transcription factor activities (TFAs), which are supposed to sum up in a single number the concentration of the transcription factor at a certain experimental point and its binding affinity to its target genes. The techniques used are modified forms of linear regression, where the TFAs are obtained as regression coefficients. These models were able to obtain results in broad accordance with the existing biological knowledge, and have the advantage of being fast and practical for genome-wide analysis. However, a major limitation of these methods is that TFAs inferred are constant across genes, i.e. they can only infer the mean influence of a transcription factor on its target genes. Also, none of these methods are probabilistic and therefore it is difficult to see how credibility intervals can be obtained, as well as how the models can be made robust against false positives (a notorious problem of ChIP data).

Other approaches based on Bayesian networks addressed these limitations, but the

additional computational complexity ruled out genome wide analysis.

Here, we propose a probabilistic model that extends linear regression in order to capture gene-specific regulatory effects. We model separately the transcription factors' active protein concentrations and the intensity with which transcription factors regulate their target genes. To reduce the number of variables involved, we choose appropriate prior distributions on the regulatory intensities and the protein concentrations. The regulatory intensities are given a zero-mean Gaussian prior, while the concentrations are modelled as a Markov chain with a different temporal continuity parameter for each transcription factor, thus allowing us to identify which transcription factors exhibit the greatest temporal coherence. Our new model has the advantage of providing probabilistic estimates for the intensities of the regulatory relationships, hence allowing the genome-wide quantitative reconstruction of the dynamical process of transcriptional regulation.

Models with a Markov chain prior on continuous valued latent variables are special cases of dynamical Bayesian networks known as state space models, or Kalman filters, and these models have previously been used in microarray time-series analysis. However, they did not make use of prior knowledge about potential regulatory interactions to explicitly infer the activity of transcription factors. This knowledge greatly reduces the search space, allowing genome-wide applications to become feasible and reducing the need for substantial experimental replication.

We validate the model both on synthetic data and real data from two yeast time series, using network architecture data obtained from ChIP experiments.

Our results are largely confirmed in the biological literature, but, using the gene-specific nature of our model, we also manage to predict biologically plausible regulatory relationships which are not documented in the literature. For example, we see that the same transcription factor can have opposite effects on different genes and, combining this with functional information about the genes and transcription factors, we obtain predictions about which genes are enhanced and which genes are repressed by a specific transcription factor. We exemplify this procedure considering some well studied transcription factors in yeast, and obtain novel biologically plausible predictions. For example, considering the targets of the well studied transcription factor ACE2, we obtain that the most strongly promoted genes are some of its known targets such as CTS1 and YER124C. However, the model also predicts a negative regulation for ACE2 acting on NCE4. This is not documented in the biological literature, but is biologically plausible as ACE2 is involved in terminating mitosis and NCE4's main function during the cell cycle is to preserve DNA stability during mitosis.

The probabilistic nature of our model also means that we can identify false positives in the ChIP data as regulatory relationships below a certain significance threshold. Specifically, by considering the posterior distribution for the gene-specific regulatory intensities, we can determine whether the mean of the posterior distribution is significantly different from zero using the posterior standard deviation.