Gaussian process models of gene expression and gene regulation

Antti Honkela

Bayesian methods are well-suited for analysis of molecular biology data as the data sets practically always consist of very few samples with a high noise level. We have studied models of gene transcription regulation based on time series gene expression data in collaboration with Neil D. Lawrence and Magnus Rattray of the University of Sheffield. This is a very challenging modelling task as the time series are very short, typically at most a dozen time points.

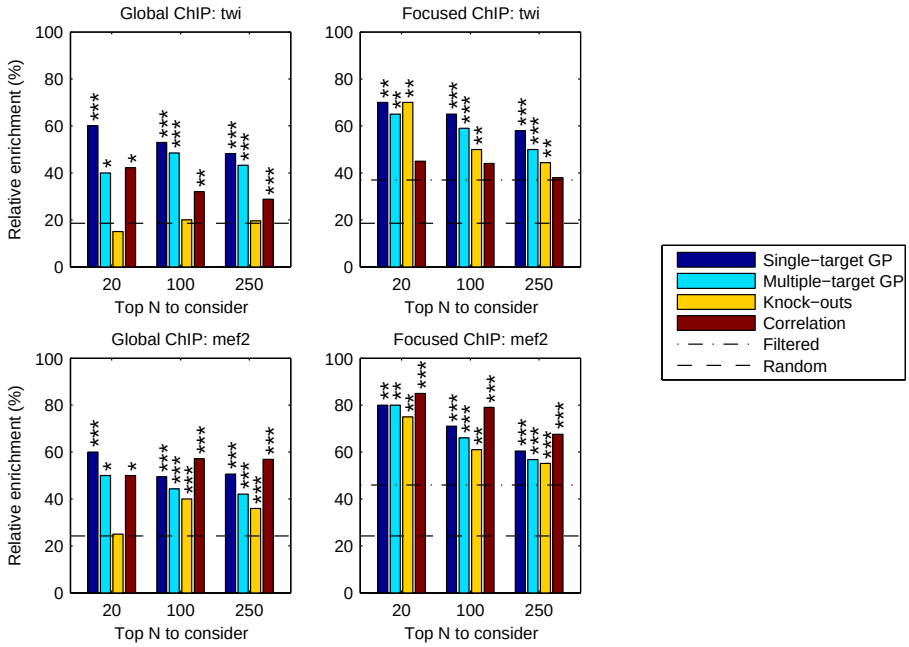

Figure 1: Evaluation results from [2] of two variants of the proposed GP-based ranking methods and two alternatives showing the relative frequency of positive predictions among N top-ranking targets ("global" evaluations) and among N top genes with annotated expression in mesoderm or muscle tissue ("focused" evaluations) for two studied transcription factors. The dashed line denotes the frequency in the full population and the dash-dot line within the population considered in focused evaluation. The bars show the frequency of targets with ChIP-chip binding within 2000 base pairs of the gene. p-values of results significantly different from random are denoted by '***': $p < 0.001$, '**': $p < 0.01$, '*': $p < 0.05$.

Extending the model of [1] of single input motif systems, i.e. where a single transcription factor regulates a number of genes, we have developed a method of ranking putative targets of transcription factors based on expression data [2]. This is achieved by imposing a Gaussian process (GP) prior on the latent continuous time transcription factor gene expression profile, which drives a linear ODE model of transcription factor protein translation and target gene transcription. This linear ODE model leads to a joint GP model for all observable gene expression values and allows exact marginalisation of the latent functions. Candidate target genes can be ranked using model likelihood.

We have applied the model to genome-wide ranking of potential target genes of transcription factors. Fig. 1 shows results from experiments with key regulators of *Drosophila* mesoderm and muscle development. They show very high accuracy in terms of enrichment of detected transcription factor binding near the predicted target genes [2]. An implementation of the method is available in Bioconductor for R [3].

References

[1] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* 24(16):i70–i75, 2008.

[2] A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. M. Furlong, N. D. Lawrence, and M. Rattray. Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A* 107(17):7793–7798, 2010.

[3] A. Honkela, P. Gao, J. Ropponen, M. Rattray, and N. D. Lawrence. tigre: Transcription factor Inference through Gaussian process Reconstruction of Expression for Bioconductor. *Bioinformatics* 27(7):1026–1027, 2011.