Supplementary material to "A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription"

Guido Sanguinetti
Magnus Rattray
Neil D. Lawrence

13th April 2006

In this document we discuss in depth some details of the computational model which could not be discussed in the main paper due to space limitations. We also present more results in terms of TFA profiles and lists of active transcription factors.

1 Dependence of the model on the choice of p-value in the ChIP data

In the paper we chose to discretize the ChIP data by considering as positive only bindings corresponding to a p-value smaller than 10^{-3}. This, as far as we are aware, is the standard procedure followed by the authors of all other papers applying regression-based methods to the problem of integrating microarray and ChIP data, so it was an obvious choice in order to compare our results with other methods.

The p-value cut-off was originally suggested in [1] as providing a low level of false positives (estimated at 5%) and an acceptable level of false negatives (the authors estimated that approximately a third of all bindings went undetected). It must be borne in mind, though, that binding is only a necessary condition for regulation [4]. Therefore, the fraction of false positives in the regulatory relationships can be expected to be much higher.

To test the robustness of our model upon changes in the cut-off p-value, we ran our model on the cell-cycle data set with three different connectivity matrices: one was obtained by using the customary cut-off at $p = 10^{-3}$, another was obtained by taking $p = 2 \times 10^{-4}$ and a third by taking $p = 5 \times 10^{-3}$. The data sets obtained were very different: the most stringent p-value led to a smaller network with 1309 genes and 88 transcription factors, while at the other end $p = 5 \times 10^{-3}$ led to a large network with 3130 genes and 112 transcription

p-value	possible links	significant links
2×10^{-4}	2111	486
1×10^{-3}	3656	716
5×10^{-3}	7200	1007

Table 1:
Possible regulations and significant regulations for three different choices of cut-off in the p-value for ChIP data

factors. All the results can be obtained by using the MATLAB code available for download at http://umber.sbs.man.ac.uk/resources/puma.

We then constructed effective connectivity matrices by retaining only regulatory relations that were predicted to be significant (at 95% significance level) by our model. The results of the analysis are summarised in Table 1.

Not surprisingly, the fraction of significant regulations (out of the possible total) is higher the more stringent the cut-off chosen. However, while this fraction is very similar in the two most stringent cases (23.0% and 19.6% respectively), it is smaller for $p = 5 \times 10^{-3}$ (approximately 14.0%). This suggests that the number of false positives at $p = 5 \times 10^{-3}$ might be too large, forcing the model to explain behaviours that are inconsistent and resulting in fewer confident predictions.

Comparing the results between the different experiments, we see that the two more stringent p-values give similar results, with approximately 70% of effective regulatory relations shared between the two predictions. This is somewhat surprising if we consider how different the two networks are. The run using $p = 5 \times 10^{-3}$ gave less consistent results, with only approximately 46% of effective regulatory relations predicted both at $p = 10^{-3}$ and $p = 5 \times 10^{-3}$. These relations were almost all present also in the run using the most stringent p-value. These results changed only slightly if we altered the significance threshold for the effective connectivity matrix.

This analysis indicates that, consistently with the suggestions of [1], $p = 10^{-3}$ provides a good choice of a cut-off to discretize ChIP data, as it seems to capture a large enough number of regulatory relationships while at the same time keeping the number of false positives at a reasonable level.

2 Global analysis of regulatory networks

We then considered the global aspects of the inferred regulations in both the cell cycle case and metabolic cycle case. To assess the significance of a relationship we considered the ratio between the changes across time of the gene-specific TFA and the associated standard deviation, considering ratios greater than 2 to be significant at 95% confidence level. The results of this analysis are summed up in Table 2.

In the cell cycle data set (with connectivity obtained using a p-value of 10^{-3}), the model inferred 716 significant regulatory relations. These were due

Data set	No of genes	No of TFs	genes with multiple regulators
Cell cycle	522	47	23%
Metabolic cycle	2167	151	11%

Table 2:
Global properties of the networks of significant regulations for the two data sets studied.

to 47 transcription factors acting on 522 genes. Out of the 47 transcription factors, 27 are confirmed transcription factors active during the cell cycle [3]. These account for 466 regulatory relations. Of the 522 genes involved, 119 had more than one significant regulator. A list of the transcription factors involved, together with the number of genes they significantly regulate and a comparison with the data from [3] is included in the attached spreadsheet CellCycleTF.xls.

In the metabolic cycle data, our model detected 2410 significant regulations involving 151 transcription factors and 2167 genes. Notice that in this data set the fraction of significant regulations out of the total possible is much higher than in the cell cycle at the same level of significance (approximately 42% versus 20%). This is probably due to the fact that, in the metabolic cycle data set, we were able to use the noise information extracted at probe level using the mmgMOS algorithm [2], resulting in a more principled treatment of the noise. In the metabolic cycle, 236 genes appear to have multiple significant regulators, five of which were regulated by three transcription factors and one by four. A list of the transcription factors involved and the number of genes each of them regulates is included in the attached spreadsheet MetabolCycleTF.xls.

3 Further TFAs

In the main paper there was only space to show the inferred TFAs only in very few cases (ACE2 for the cell cycle data set, LEU3 and ACE2 for the metabolic cycle data set). We show here more gene-specific TFAs and compare them with non-specific profiles obtained from regression. Further examples can be obtained by using the online MATLAB code.

Figure 2 shows the TFAs of four more transcription factors that are involved in the cell cycle according to our results. The TFA obtained by regression for these transcription factors is shown in the first column, while the other columns show the gene-specific TFAs obtained with our model for the three most significantly regulated targets of these transcription factors. Notice that some gene-specific TFAs look quite different from the TFAs obtained by regression. For example, the TFA obtained by regression for STE12 seem to be dominated by white noise, while the gene specific TFA on the most significant targets shows a very different behaviour, being stationary for the first part of the cycle and peaking towards the end. The gene-specific TFAs of MBP1's three main targets again show different behaviours among them, and in turn different from the regression picture.

Figure 1 shows the TFAs of four more transcription factors that significantly regulate five or more genes in the metabolic cycle but do not have periodic expression according to [5]. These are ABF1, ARO80, FHL1 and SMP1. The TFA obtained by regression for these transcription factors is shown in the first column, while the other columns show the gene-specific TFAs obtained with our model for the three most significantly regulated targets of these transcription factors. These results indicate that, even if the expression levels of these transcription factors is not periodic, their gene-specific activities are to be considered periodic for many genes. We would suggest that the appropriate criterion to determine whether a transcription factor is involved in a periodic cellular process is whether its activities, rather than its expression level, display periodic behaviour.

References

[1] Tong Ihn Lee, Nicola J. Rinaldi, Francois Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Thompson, Itamar Simon, Julia Zeitlinger, Ezra G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Ernest Fraenkel, David K. Gifford, and Richard A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[2] Xuejun Liu, Marta Milo, Neil D. Lawrence, and Magnus Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.

[3] Nicholas M. Luscombe, M. Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A. Teichmann, and Mark Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[4] Rebecca Martone, Ghia Euskirchen, Paul Bertone, Stephen Hartman, Thomas E. Royce, Nicholas M. Luscombe, John L. Rinn, F. Kenneth Nelson, Perry Miller, Mark Gerstein, Sherman Weissman, and Michael Snyder. Distribution of nf-κb-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences USA*, 100(21):12247–12252, 2003.

[5] Benjamin P. Tu, Andrzej Kudlicki, Maga Rowicka, and Steven L.McKnight. Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5715):1152–1158, 2005.

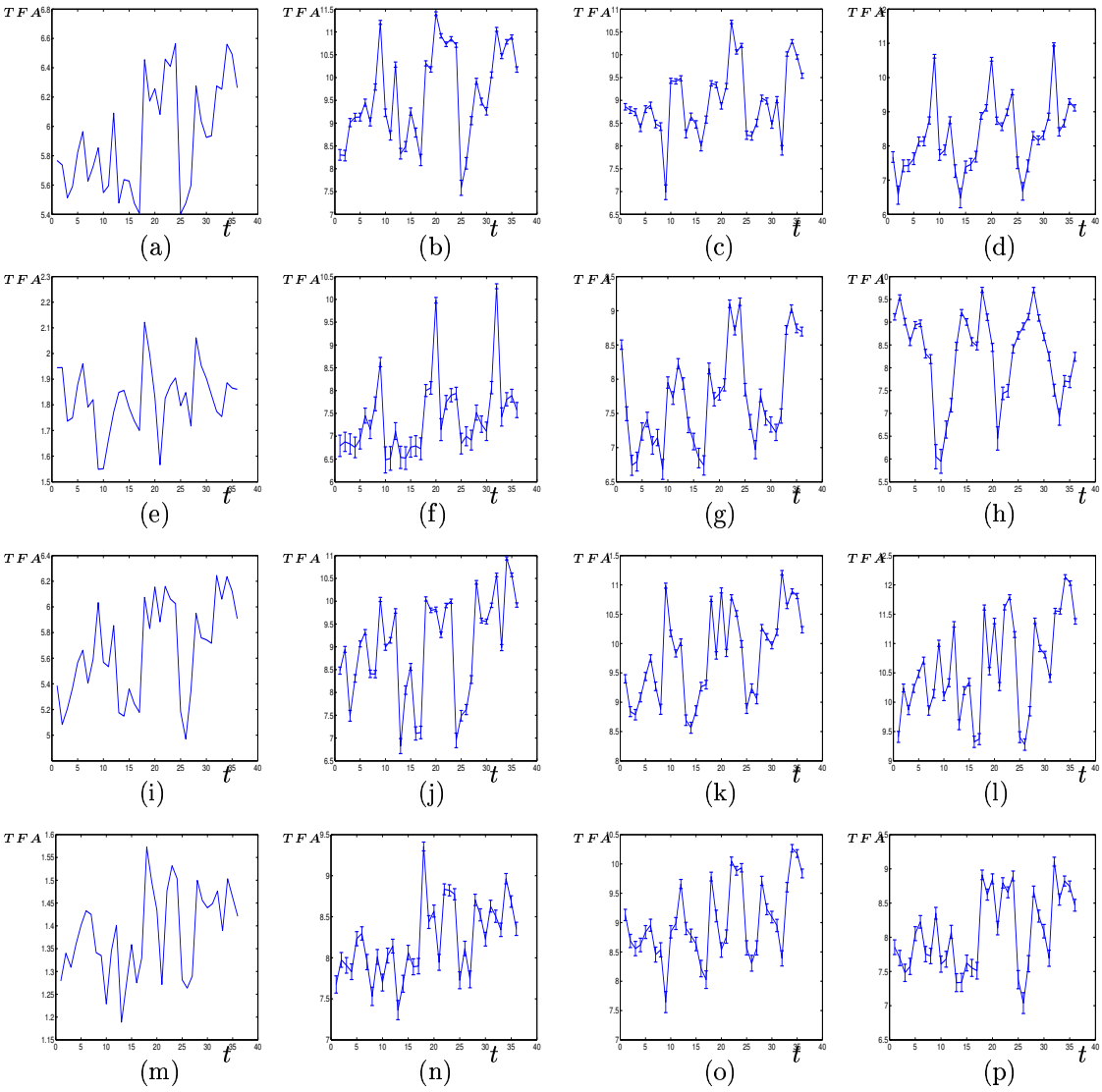

Figure 1: TFAs and gene-specific TFAs for some transcription factors active in the yeast metabolic cycle which have non periodic expression levels. (a) TFA of ABF1 obtained by regression. (b-d) TFA of ABF1 for its three main targets, YOR309C, RFA2 and YPL012W respectively. (e) TFA of ARO80 obtained by regression. (f-h) TFA of ARO80 for its three main targets, YNL124W, ARH1 and ARO10 respectively. (i) TFA of FHL1 obtained by regression. (j-l) TFA of FHL1 for its three main targets, RPL9A, RPL14B and RPL9B respectively. (m) TFA of SMP1 obtained by regression. (n-p) TFA of SMP1 for its three main targets, SLT2, GRX5 and HKR1 respectively.

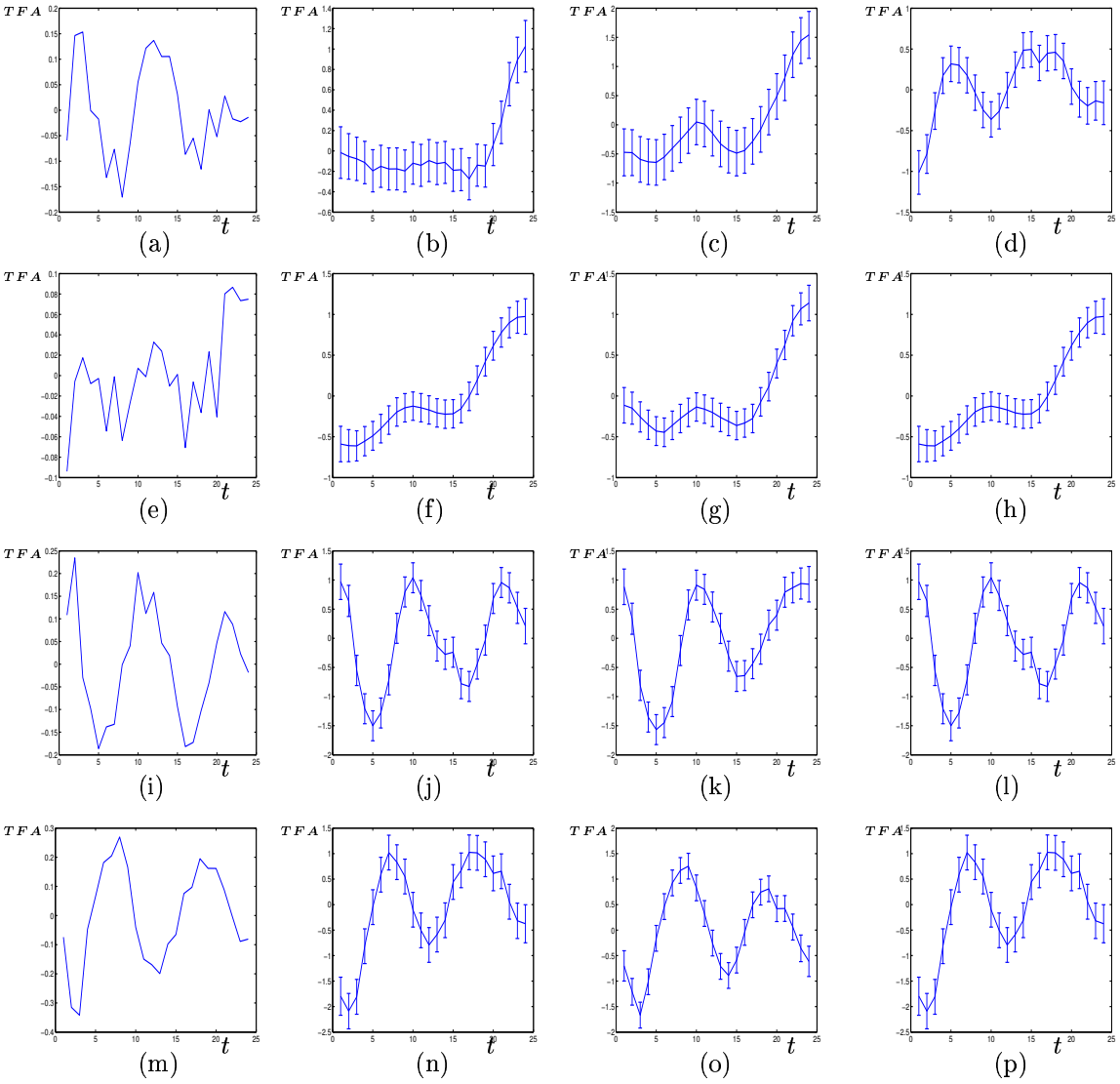

Figure 2: TFAs and gene-specific TFAs for some transcription factors active in the yeast cell cycle. (a) TFA of NDD1 obtained by regression. (b-d) TFA of NDD1 for its three main targets, PHO3, YDR033W and NCE102 respectively. (e) TFA of SWI5 obtained by regression. (f-h) TFA of SWI5 for its three main targets, PIR1, PIR3 and ASH1 respectively. (i) TFA of STE12 obtained by regression. (j-l) TFA of STE12 for its three main targets, FUS1, KAR4 and SST2 respectively. (m) TFA of MBP1 obtained by regression. (n-p) TFA of MBP1 for its three main targets, MRP8, AGA1 and YMR215W respectively.