MODELLING UNCERTAINTY IN TRANSCRIPTOME MEASUREMENTS ENHANCES NETWORK COMPONENT ANALYSIS OF YEAST METABOLIC CYCLE

C.Q. Chang, Y.S. Hung

Department of Electrical and Electronic Engineering,
University of Hong Kong,
Pokfulam Road, Hong Kong.
cqchang / yshung @ee.hku.hk

M. Niranjan *

School of Electronics and Computer Science,
University of Southampton,
Highfield, Southampton, UK.
mn@ecs.soton.ac.uk

ABSTRACT

Using high throughput DNA binding data for transcription factors and DNA microarray time course data, we constructed four transcription regulatory networks and analysed them using a novel extension to the network component analysis (NCA) approach. We incorporated probe level uncertainties in gene expression measurements into the NCA analysis by the application of probabilistic principal component analysis (PPCA), and applied the method to data from yeast metabolic cycle. Analysis shows statistically significant enhancement to periodicity in a large fraction of the transcription factor activities inferred from the model. For several of these we found literature evidence of post-transcriptional regulation. Accounting for probe level uncertainty of microarray measurements leads to improved network component analysis. Transcription factor profiles showing greater periodicity at their activity levels, rather than at the corresponding mRNA levels, for over half the regulators in the networks points to extensive post-transcriptional regulations.

Index Terms— Network component analysis, Transcription regulation, Microarray.

1. INTRODUCTION

Given time course data on the transcriptome of an organism, the modelling task is to factorize the expression matrix into temporal profiles of activities of the regulators and weights corresponding to sensitivities between target genes and transcription factors. This factorization is written as a matrix equation: $\mathbf{X} = \mathbf{A}\,\mathbf{S}$, where $\mathbf{X}$ is the gene expression matrix of N genes times K time points, $\mathbf{A}$ denotes the regulatory binding matrix of dimension N genes times M transcription factors and $\mathbf{S}$, the activity profiles of the transcription factors, M times K. The binding matrix $\mathbf{A}$ is derived from so called ChIP-chip experiments [1].

The technique of Network Component Analysis (NCA) was developed by Liao *et al.* [2] to achieve the factorization above. They show that the connectivity matrix $\mathbf{A}$ has to satisfy a series of conditions to achieve unique factorization. By careful construction of subnetworks from the binding data, a subset of $\mathbf{X}$ can be factorized by least squares fitting. Chang *et al.* [3, 4], extend the NCA model by a sequential subspace projection approach and propose faster versions of the decomposition. An alternative formulation for factorizing the gene expression matrix was developed by Sanguinetti *et al.* [5]. An important aspect of any inference procedure that is based on experimental data is the quantification of measurement uncertainties, and the propagation of the effect of uncertainties in downstream inference. In this context, Milo *et al.* [6] developed a procedure to quantify probe level uncertainty in *Affymetrix* microarrays by fitting a *Gamma* density function. Uncertainties quantified in this manner can be propagated through downstream analysis [7].

In this paper we demonstrate how accounting for probe level uncertainties in regulatory network inference enhances the analysis. We use time course microarray data from Tu *et al.* [8], and estimate four regulatory subnetworks, constructed to satisfy identifiability constraints derived by Liao *et al.* [2].

2. APPROACH

2.1. Network component analysis

Let $\mathbf{X}$ ($N \times K$) be the microarray measurement of the expression of N genes over K time points, $\mathbf{S}$ ($M \times K$) be the activities of the associated M transcription factors (TF) over the same time span, $\mathbf{A}$ ($N \times M$) be the regulation strength of the transcription factor activities (TFA) on the gene expressions, and $\mathbf{\Gamma}$ be the measurement noise. The following linear model can be used to describe the gene regulatory network [2]:

$$\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma} \tag{1}$$

The technique of network component analysis (NCA) was developed in [2] to estimate the unknown connectivity matrix $\mathbf{A}$ and TFA matrix $\mathbf{S}$ if the following three constraints (referred to as *NCA criteria*) are satisfied: **(a)** the connectivity matrix $\mathbf{A}$ is of full column rank; **(b)** when any one column of $\mathbf{A}$ is removed together with the rows of $\mathbf{A}$ where the corresponding entries of the removed column is nonzero, the re-

*This work was initiated during a visit by the third author to the University of Hong Kong as a Sir William Mong Fellow

sulting sub-matrix of $\mathbf{A}$ is still of full-column rank; and **(c)** $\mathbf{S}$ has full row rank. NCA estimates $\mathbf{A}$ and $\mathbf{S}$ by minimizing the following objective function through alternating least squares (ALS),

$$\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|^2 \quad s.t. \quad \mathbf{A} \in \mathbf{Z}_0, \tag{2}$$

where $\mathbf{Z}_0$ defines the network topology constraint for $\mathbf{A}$. The difficulty with NCA is that the iterative solution is computationally demanding and the ALS does not always converge to the global minimum.

In subsequent work, Eq. 2 was regularized to improve the stability of the estimation in an algorithm we will refer to as NCAr, using the objective function:

$$\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|^2 + \lambda\|\mathbf{S}\|^2 \quad s.t. \quad \mathbf{A} \in \mathbf{Z}_0. \tag{3}$$

A more computationally efficient method, called FastNCA, was proposed by Chang *et al.* [3, 4] to estimate $\mathbf{A}$ and $\mathbf{S}$ through fitting the model by a series of subspace projections, using an analytical solution of the problem in the noiseless case.

2.2. Modelling uncertainty in microarray data

We note that the information contained in the measured data for the model of Eq. 1 comes in two different forms, namely, (i) the noise statistics D_n (see below) of the measurements. and (ii) the microarray gene expression measurements X (after noise reduction). In accordance with these two sources of information, which are of different nature, we propose a two-step procedure, i.e. probabilistic PCA, for noise removal, followed by FastNCA, to make use of cleaned microarray data to recover the connectivity matrix of the network.

We assume that in Eq 1 the transcription factor activity, the noise, and thus the gene expression are all Gaussian distributed, $\mathbf{s}_n \sim N(\tilde{\boldsymbol{\mu}}, \mathbf{I})$ and $\boldsymbol{\gamma}_n \sim N(0, \sigma^2\mathbf{I})$, where $\mathbf{x}_n, \mathbf{s}_n$ and $\boldsymbol{\gamma}_n$ are the nth column of $\mathbf{X}, \mathbf{S}$ and $\boldsymbol{\Gamma}$, respectively. Let $\mathbf{s}_n = \tilde{\mathbf{s}}_n + \tilde{\boldsymbol{\mu}}$ so that $\tilde{\mathbf{s}}_n \sim N(0, \mathbf{I})$, then from Eq 1 we have

$$\mathbf{x}_n = \mathbf{A}\tilde{\mathbf{s}}_n + \boldsymbol{\mu} + \boldsymbol{\gamma}_n, \tag{4}$$

where $\boldsymbol{\mu} = \mathbf{A}\tilde{\boldsymbol{\mu}}$. This model is used in the development of probabilistic principal component analysis.

Microarray probe level uncertainty is expressed explicitly in the model of Eq. 4, so that the measurement with uncertainty is

$$\hat{\mathbf{x}}_n = \mathbf{x}_n + \boldsymbol{\nu}_n = \mathbf{A}\tilde{\mathbf{s}}_n + \boldsymbol{\mu} + \boldsymbol{\gamma}_n + \boldsymbol{\nu}_n, \tag{5}$$

where the uncertainty vector $\boldsymbol{\nu}_n \sim N(0, \mathbf{D}_n)$ is uncorrelated Gaussian with $\mathbf{D}_n$ diagonal and assumed known in this model. In this paper, $\mathbf{D}_n$ will be estimated from probe-level analysis of the gene expression data. Based on such a model specification, we can apply the EM algorithm developed in [9] to obtain maximum likelihood estimates of the unknown $\mathbf{A}$, $\tilde{\mathbf{s}}_n$, $\boldsymbol{\mu}$, and σ^2. With the maximum likelihood estimates of $\mathbf{A}, \tilde{\mathbf{s}}_n$ and $\boldsymbol{\mu}$, denoted as $\bar{\mathbf{A}}, \bar{\mathbf{s}}_n$ and $\bar{\boldsymbol{\mu}}$, respectively, obtained from the extended probabilistic PCA model of Eq. 5, we can estimate $\mathbf{x}_n$ as

$$\bar{\mathbf{x}}_n = \bar{\mathbf{A}}\bar{\mathbf{s}}_n + \bar{\boldsymbol{\mu}}. \tag{6}$$

We now apply factorizations NCAr and FastNCA to the reconstructed matrix $\bar{\mathbf{X}}$ whose columns are the estimated $\bar{\mathbf{x}}_n$ in Eq. 6. Since the uncertainty has been accounted for in the extended probabilistic PCA model, applying NCA to $\bar{\mathbf{X}}$ instead of $\mathbf{X}$ directly is expected to give improved performance. We refer this approach to accounting for uncertainty in NCA as uNCAr and uFastNCA, corresponding to NCAr and FastNCA, respectively.

Networks to which NCA is applied typically have much larger number of genes than the number of time points. Hence, to have better computational efficiency and statistical stability, in the preprocessing step of probabilistic PCA we work with the transpose of the gene expression matrix with the following model induced from the original NCA model of Eq. 1,

$$\mathbf{X}^T = \mathbf{S}^T\mathbf{A}^T + \boldsymbol{\Gamma}^T. \tag{7}$$

3. RESULTS

3.1. Performance of uNCA on synthetic data

In an initial study (Fig. 1), we compared the performance of uNCA and conventional NCA across different levels of uncertainty on synthetic data. The average levels of uncertainty are set at 1, 2, 5, and 10, which, in signal processing terms approximates, $0\,dB$, $3\,dB$, $7\,dB$, and $10\,dB$, respectively. Note that for the microarray data of yeast metabolic cycle [8] to be studied in this paper, the uncertainty level is about $3.5\,dB$. For each case, 100 Monte Carlo simulations were performed to calculate the average performance in terms of mean square error of the estimated transcription factor activities as compared to ground truth. In this simulation we use FastNCA and its counterpart under measurement uncertainty, since Liao's NCA is too slow to perform these simulations in reasonable time. However, the conclusion should also apply to Liao's NCA as the model specifications and objective functions are the same. It was found that for every case and every Monte Carlo simulation uNCA performs consistently better than conventional NCA.

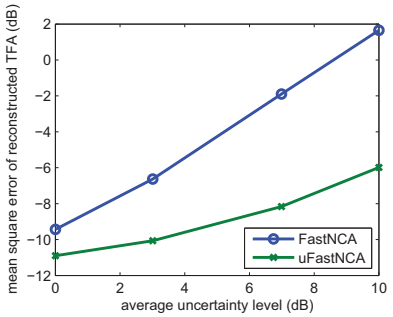

Figure 1 Simulation results on synthetic data

3.2. Probe-level processing of the microarray data

In the data analysis, we first estimate the measurement uncertainty from analysis of probe-level data by multi-mgMOS [10], then account for the uncertainty in the extended probabilistic principal component analysis to reduce the uncertainty and get a less-noisy reconstruction of the gene expression matrix, and at last apply NCA methods (NCAr and FastNCA) to the reconstructed gene expression matrix to get estimates of the connectivity matrix $\mathbf{A}$ and also the transcription factor activity (TFA) profiles $\mathbf{S}$.

3.3. Sub-network design

Of over 200 yeast transcription factors annotated in databases, there are 47 cell cycle related transcription factors and 83 metabolism related transcription factors. Of these, 29 and 49 were found in Lee *et al.*'s ChIP-chip experiment [1], respectively, and were used in the construction of subnetworks for further analysis. As in previous work, we set a threshold p-value of 0.001, to obtain the connectivity topology [2, 4].

Four sub-networks were built to study these cell cycle and metabolism related transcription factors. All these 4 subnetworks satisty the conditions required by network component analysis [2, 4].

- Network 1: as in Liao's paper [2], it contains 33 transcription factors, out of which 11 are related to cell cycle. In the network there are 1300 genes regulated by and only by these 33 transcription factors.
- Network 2: it contains 27 cell cycle related transcription factors, and 802 genes are regulated by and only by these 27 cell cycle related transcription factors. The remaining 2 cell cycle related transcription factors, namely `ASH1` and `RIM101`, are excluded from the network in order to satisfy the conditions required by network component analysis [2, 4].
- Network 3: it contains 24 metabolism related transcription factors and 512 genes regulated by and only by these 24 transcription factors.
- Network 4: it contains another set of 24 metabolism related transcription factors and 192 genes. Network 3 and network 4 cover 48 metabolism related transcription factors, and the remaining one, namely `MAL13`, is not included in order to satisfy the conditions required by network component analysis.

Analysing the inferred temporal profiles of the transcription factors in Network 1, we found that for 18 of the 33 regulators, periodicity is enhanced at the estimated activity level over the periodicity at the corresponding mRNA level: `ABF1`, `ACE2`, `FHL1`, `FKH1`, `GCN4`, `GRF10`, `HIR1`, `HIR2`, `MBP1`, `MCM1`, `NDD1`, `NRG1`, `RLM1`, `SKN7`, `SMP1`, `STB1`, `SWI4` and `SWI6`. Most of the periodicity enhancement is visible from the plots of rows of matrix $\boldsymbol{S}$, but we confirmed this by computing Fourier transforms. Amongst these 18 is the leucine zipper protein, `GCN4`, implicated in amino acid biosynthesis. This protein is known to be post-translationally regulated [11], thus its mRNA levels are not clear indicators of its dynamical regulatory action. Similarly, `FKH1`, `HIR1` and `NDD1` are targets of the `Puf` family of RNA binding proteins (RNP) as demonstrated in the genome-wide experiments published recently [12]. Association with RNPs is suggestive of post transcriptional regulation of these transcription factors. `SWI6` is another gene whose periodicity is enhanced at the protein level. This protein is regulated in a complex manner [13] by periodic movement in and out of the nucleus. It is suggested that the protein is synthesized in an inactive form and is probably activated only after shuttling in and out of the cytoplasm. The active form of the protein being realised after such a post-translational process is supportive of the observation of increased periodicity. Such difference between periodicity at the two different levels of observation has also been noted by Sanguinetti *et al.* [5], using their Bayesian state-space model, showing that 41 transcription factors they identified as regulating metabolic cycle in Tu *et al.*[8]'s data, do not show periodic expression at mRNA level. Similarly in the other three networks, we noted several transcription factors ($15/27$, $13/24$ and $11/24$ for networks two, three and four) for which periodicity of the inferred activity profile was significantly higher than that of the mRNA profile.

3.4. Modelling uncertainty enhances periodicity estimates

For each of the four network, we used the R package `GeneCycle` to test the periodicity of the mRNA profile and the periodicity of the TFA profile estimated by NCA methods. The periodicity is quantified by a p-value from Fisher's g test, where a smaller p-value means greater periodicity. Simulations showed that in general the TFA of a transcription factor is more periodic than its mRNA profile, and the TFAs estimated by uNCAr and uFastCNA that account for measurement uncertainty are more periodic than those estimated by NCAr and FastNCA that do not account for the measurement uncertainty. This was confirmed in a formal test for statistical significance, as described below.

The TFs in the four networks are tested. In total there are 108 TFs. In the following table, the first column "A > B" means TFA profiles estimated by method A are more periodic than those from method B; for the case of "mRNA", TFA is compared with the corresponding raw mRNA expression level; the second column is the number of TFs for which A > B; and the third column is the p-value of the test of "A > B" against the null-case "A is not different from B". In the null-case the number of TFs for which "A > B" has a binomial distribution of $n = 108$ and $p = 0.5$, i.e., $B(108, 0.5)$. The p-values are calculated from this binomial distribution.

	# of TFs	p-value
NCAr > mRNA	71	$3.42E-4$
FastNCA > mRNA	55	0.39
uNCAr > mRNA	75	$1.38E-5$
uFastNCA > mRNA	75	$1.38E-5$
uNCAr > NCAr	77	$2.16E-6$
uFastNCA > FastNCA	73	$7.45E-5$

Table 1: Test of enhancement of periodicity.

The test of significance above confirms that: (a) Compared to the mRNA profile, the TFA activities estimated by various NCA algorithms are statistically more periodic, except in the case of the conventional FastNCA algorithm that does not account for uncertainty; and (b) the TFA profiles estimated by uNCAr and uFastNCA which account for measurement uncertainty are more periodic than that of NCAr and FastNCA which are conventional NCA algorithms not accounting for the uncertainty. We recall similar enhancements in reconstruction error demonstrated on synthetic data, when measurement uncertainty was built into the model.

4. CONCLUSIONS

We have constructed four subnetworks that satisfy the constraints of Network Component Analysis and applied the inference procedure to yeast metabolic cycle data. We have used probabilistic principal component analysis to propagate probe level uncertainty in microarray measurements through such network analysis. Our study shows clear enhancement of periodicity, for a number of transcription factors, at the inferred regulator activity level when compared to the periodicity at the mRNA expression level. This is reasonably attributed to post transcriptional regulation of the transcription factors. We have found several examples in the literature for evidence of post transcriptional and post translational regulation that corroborates this view.

5. REFERENCES

[1] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, 2002, 45.

[2] J C Liao, R Boscolo, Y L Yang, L M Tran, C Sabatti, and V P Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *PNAS*, vol. 100, no. 26, pp. 15522–15527, 2003.

[3] C. Q. Chang, Y. S. Hung, P. C. W. Fung, and Z. Ding, "Network component analysis for blind source separation," in *Proc. 2006 International Conference on Communications, Circuits and Systems (ICCCAS'2006)*, Guilin, China, June 2006, vol. 1, pp. 323–326.

[4] C. Chang, Z. Ding, Y. S. Hung, and P. C. W. Fung, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics*, vol. 24, no. 11, pp. 1349 – 1358, 2008.

[5] G. Sanguinetti, M. Rattray, and N. D. Lawrence, "A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription," *Bioinformatics*, vol. 22, no. 14, pp. 1753 – 1759, 2006.

[6] M. Milo, A. Fazeli, M. Niranjan, and N.D. Lawrence, "A probabilistic model for the extraction of expression levels from oligonucleotide arrays," *Biochem. Soc. Trans*, vol. 31, pp. 1510–1512, 2003.

[7] M. Rattray, X. Liu, G. Sanguinetti, M. Milo, and N. D. Lawrence, "Propagating uncertainty in Microarray data analysis," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 37 – 47, 2006.

[8] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight, "Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes.," *Science*, vol. 310, no. 5751, pp. 1152 – 1158, 2005.

[9] G. Sanguinetti, M. Milo, M. Rattray, and N. D. Lawrence, "Accounting for probe-level noise in principal component analysis of microarray data.," *Bioinformatics*, vol. 21, no. 19, pp. 3748 – 3754, 2005.

[10] X Liu, M Milo, ND Lawrence, and M Rattray, "A tractable probabilistic model for affymetrix probe-level analysis across multiple chips," *Bioinformatics*, vol. 21, no. 18, pp. 3637–3644, 2005.

[11] P. P. Mueller, S. Harashima, and A. G. Hinnebusch, "A segment of GCN4 mRNA containing the upstream AUG codons confers translational control upon a heterologous yeast transcript," *PNAS*, vol. 84, no. 9, pp. 2863–2867, 1987.

[12] A. P. Gerber, D. Herschlag, and P. O. Brown, "Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast," *PLoS Biol.*, vol. 2, no. 3, pp. e79, 2004.

[13] E. Queralt and J. C. Igual, "Cell cycle activation of the `Swi6p` transcription factor is linked to nucleocytoplasmic shuttling," *Molecular and Cellular Biology*, vol. 23, no. 9, pp. 3126 – 3140, 2003.