Identifying submodules of cellular regulatory networks.

Guido Sanguinetti[1], Magnus Rattray[2] and Neil D. Lawrence[1]

[1] Department of Computer Science, University of Sheffield
211 Portobello Street, Sheffield S1 4DP, U.K. {guido, neil}@dcs.shef.ac.uk
[2] School of Computer Science, University of Manchester
Oxford Road, Manchester M13 9PM, U.K. magnus@cs.man.ac.uk

Abstract. Recent high throughput techniques in molecular biology have brought about the possibility of directly identifying the architecture of regulatory networks on a genome-wide scale. However, the computational task of estimating fine-grained models on a genome-wide scale is daunting. Therefore, it is of great importance to be able to reliably identify submodules of the network that can be effectively modelled as independent subunits. In this paper we present a procedure to obtain submodules of a cellular network by using information from gene-expression measurements. We integrate network architecture data with genome-wide gene expression measurements in order to determine which regulatory relations are actually confirmed by the expression data. We then use this information to obtain non-trivial submodules of the regulatory network using two distinct algorithms, a naive exhaustive algorithm and a spectral algorithm based on the eigendecomposition of an affinity matrix. We test our method on two yeast biological data sets, using regulatory information obtained from chromatin immunoprecipitation.

1 Introduction

The modelling of cellular networks has undergone a revolution in recent years. The advent of high throughput techniques such as microarrays and chromatin immunoprecipitation (ChIP [1, 2]) has resulted in a rapid increase in the amount of data available, so that it is possible to measure on a genome-wide scale both the expression levels of thousands of genes and the architecture (connectivity) of the regulatory network which links genes to their regulators (transcription factors). However, this data is often very noisy, and the sheer amount of data makes the development of quantitative fine grained models impossible.

Gene networks are frequently modelled in very different ways at different scales [3]. Network modelling at the genome-wide scale is often limited to the topology of networks. For example, Luscombe *et al.* used a large database constructed by integrating all available data on transcriptional regulation from a variety of sources (ChIP-on-chip, protein interaction data, *etc.*) to model the changes in the topology of the yeast regulatory network in different experimental conditions [4]. While this result was *per se* of great importance in furthering our

understanding of transcriptional regulation, it is not clear how this approach could be used to model the *dynamics* of the system. At the other end of the spectrum [5], small networks consisting of a few transcription factors and their established target genes are often modelled in a realistic fine grained way, allowing for a quantitative explanation of qualitative behaviours in the cellular processes such as cycles, spatial gradients, *etc.*

While these fine grained models are often very successful in describing specific processes, they rely on rather strong assumptions. First of all, they need the regulatory links they exploit to be *true* regulations. While there is a growing number of experimentally validated regulatory relations in a number of organisms, the main techniques to study regulatory networks on a genome-wide scale are ChIP-on-chip [1] and motif conservation [6]. However, it is well known that ChIP-on-chip only measures the binding of a transcription factor to the promoter region of the gene. While binding is obviously a necessary condition for transcription to be initiated, there is abundant biological evidence [7] that shows that it is not a sufficient condition. Therefore, we may expect that interpreting ChIP-on-chip data as evidence for regulation may lead to many false positives, which would obviously be a big problem for any fine grained model. As for motif conservation, it is often difficult to assign a motif to a unique transcription factor and large numbers of false positives can be expected. Secondly, the system modelled should be reasonably isolated from the rest of the cell. Often collateral processes are simply modelled as noise in fine grained models, and this approximation would clearly break down in the presence of strong interactions with variables not included in the model.

We recently presented a probabilistic dynamical model which allowed us to infer both the active transcription factor protein concentrations and the intensity of the regulatory links between transcription factors and their target genes [8, 9]. The model was computationally efficient so that the network could be modelled at the genome level, and its probabilistic nature meant that we could estimate the whole *probability distribution* of the concentrations and regulatory intensities, rather than just providing point estimates. This means that the significance level of the regulatory interactions could be assessed. This information can be used in many ways: for example, one may use it to obtain a refinement of the ChIP data, so that regulatory relations below a certain significance threshold are effectively treated as false positives. However, the information about the absolute value of the regulatory intensity is also of interest, since low intensity regulations (however significant) could be ignored when trying to obtain submodules of manageable size.

The main novelty of this paper is to present two algorithms to obtain submodules of regulatory networks. The first algorithm is a simple exhaustive search algorithm. While in principle this is applicable to any network with binary connectivity, it obtains biologically relevant submodules when applied to a network comprising significant regulations only. The second algorithm is a spectral method based on an eigenvalue decomposition of an affinity matrix and on a generalisation of the spectral clustering algorithm described in [10]. This takes

into account the absolute value of the regulatory intensity and has the advantage of providing a natural way of ranking the submodules according to their importance in the global cellular network.

The paper is organised as follows: we first briefly review the probabilistic model used to infer the regulatory intensities. We then present the two algorithms to identify submodules of the regulatory network. In the results section we demonstrate our approach on two yeast data sets, the benchmark cell cycle data set of [11] and the more recent metabolic cycle data set of [12]. Finally, we discuss the relative merits of the two algorithms we proposed and their validity as an alternative approach to existing graph clustering algorithms.

2 Quantitative inference of regulatory networks

Here we briefly review the probabilistic dynamical model for inference of regulatory networks proposed in [9]. This builds on the model presented in [8], which in turn extends the linear regression approach, first introduced in [13], to take into account gene-specific effects. We have (log transformed) expression level measurements y_{nt} for N genes at T time points. We assume that the binding of q transcription factors to the N genes is known (for example *via* ChIP-on-chip experiments), so that we have a binary matrix X whose nm entry X_{nm} is one if gene n is bound by transcription factor m and zero otherwise. We can then write down our model as

$$y_{nt} = \sum_{m=1}^{q} X_{nm} b_{nm} c_{mt} + \mu_n + \epsilon_{nt}. \tag{1}$$

Here b_{nm} represents the regulatory intensity with which transcription factor m enhances gene n (negative intensity models repression), c_{mt} models the (log) active protein concentration of transcription factor m at time t, μ_n is the baseline expression level of gene n and $\epsilon_{nt} \sim \mathcal{N}\left(0, \sigma^2\right)$ is an error term.

The model is then specified by a choice of prior distributions on the random variables b_{nm}, c_{mt} and μ_n. We assign spherical Gaussian priors to the regulatory intensities and the baseline expression level

$$b_{nm} \sim \mathcal{N}\left(0, \alpha^2\right)$$

$$\mu_n \sim \mathcal{N}\left(\tau, \beta\right).$$

The choice of prior distribution on the concentrations c_{mt} depends on the specific biological situation we wish to model. For example, for independent samples we may assume that the prior distribution on c_{mt} factorises along time t. As we are going to model time series data, an appropriate choice for the prior distribution on c_{mt} is a time-stationary Markov chain

$$c_{mt} = \gamma_m c_{m(t-1)} + \eta_{mt}$$

$$\eta_{mt} \sim \mathcal{N}\left(0, 1 - \gamma_m^2\right) \tag{2}$$

$$c_{m1} \sim \mathcal{N}(0,1)\,.$$

The variance in (2) is chosen so that the process is stationary, *i.e.* the expected changes over a period of time Δt depend only on the length of the time interval, not on its starting or finishing point. The parameters $\gamma_m \in [0,1]$ model the temporal continuity of the sequence c_{mt}. Values of γ_m close to 1 lead to smoothly varying samples, with contiguous time points having very similar values of concentration. On the other hand, low values of γ_m lead to samples with little correlation among time points, so that in the limit of $\gamma_m = 0$ the modelling situation of independent time points is recovered.

Having selected prior distributions for the latent variables b_{nm}, c_{mt} and μ_n we can use equation (1) to compute a joint likelihood for all the latent and observed variables

$$\begin{aligned} &p(y_{nt}, b_{nm}, c_{mt}, \mu_n | X) = \\ &= p(y_{nt}|b_{nm}, c_{mt}, \mu_n, X)\, p(b_{nm}|\alpha)\, p(c_{mt}|\gamma_m)\, p(\mu_n|\tau,\beta)\,. \end{aligned} \tag{3}$$

We can then estimate the hyperparameters α, γ_m, σ, τ and β by type II maximum likelihood. Unfortunately, exact marginalisation of equation (3) is not possible and we have to resort to approximate numerical methods. This can be done *e.g.* using a variational EM algorithm as proposed in [9], where details of the implementation are given.

Once the hyperparameters have been estimated, we can obtain the posterior distribution for the latent variables given the data using Bayes' theorem

$$p(b,c,\mu|y) = \frac{p(y|b,c,\mu)\, p(b,c,\mu)}{\int p(y,b,c,\mu)\, dbdcd\mu}. \tag{4}$$

3 Identifying submodules

3.1 Naive approach

Given the posterior probability on the regulatory intensities b_{nm}, one can associate a significance level to each regulatory interaction by considering the ratio between the posterior means and the associated standard deviations. One can then obtain a refined network structure comprising only of significant regulatory relations by considering only relations above a certain significance threshold (which can be viewed as the only parameter in this algorithm). It is then straightforward to find submodules in a regulatory network with binary connectivity. One can start with any transcription factor and subsequently include other transcription factors which have common targets with the first one. This can be iterated and it will obviously converge to a unique set of submodules. This procedure is schematically described in Algorithm 1.

3.2 Introducing the regulatory intensities

The main drawback of the procedure outlined in Algorithm 1 is that it does not take into account the information about the regulatory intensities, apart from

Algorithm 1 Identify submodules of a network with binary connectivity

Input data: set R of regulators, set G of genes, regulatory intensities b_{nm};
Construct a binary connectivity matrix X by thresholding the intensities
repeat
  Choose a regulator $r_1 \in R$. Include the set of all its target genes $G_{r1} \subset G$;
  repeat
    Include the set of regulators other than r_1 regulating genes in G_{r1}, $R_{G_{r1}} \subset R$;
    Include all genes regulated by $R_{G_{r1}}$ not included in G_{r1};
  until No new genes are found;
  Output reduced sets R_m, G_m for the submodule and $\bar{R}, \bar{G}$ for the elements not included in the submodule;
until $\bar{R}, \bar{G}$ are the empty set.

using it as a guideline to introduce thresholds of significance. Specifically, it only exploits the outputs of the probabilistic model in order to obtain a refinement of the network architecture, which is only a minimal part of the information contained in the posterior distribution over b_{nm}.

However, when trying to identify submodules considering all the available information on the regulatory intensities, we may find that there are few truly independent submodules, and it might be hard to manually determine which submodules are approximately independent. In practice, we would like to be able to have an automated way to obtain submodules.

Since our probabilistic model reconstructs transcription factors concentrations and regulatory intensities from time-course microarray data, we can interpret the regulatory strengths as a measure of the involvement of a transcription factor in the cellular processes in which its target genes participate. A standard technique for retrieving genes associated with (approximately independent) cellular processes is PCA (also known as SVD, [14]). However, the *eigengenes* retrieved by PCA are not necessarily disjoint in terms of gene participation, in particular the same genes can be represented in different eigengenes, mirroring the biological fact that the same genes can participate in more than one cellular process. While this constitutes an important piece of information in its own right, it could be a drawback from the point of view of identifying independent submodules. We therefore propose a modified algorithm which extends the spectral clustering algorithm developed in [10].

Given the posterior distribution over the regulatory intensities

$$p\left(b_{nm}|y\right) \sim \mathcal{N}\left(b_{nm}|\bar{b}_{nm}, \sigma^2_{b_{nm}}\right)$$

we construct an affinity matrix C between transcription factors using the formula

$$C_{ij} = \left|\langle \mathbf{b}_i^T \rangle\right| \left|\langle \mathbf{b}_j \rangle\right| . \tag{5}$$

Here, $\langle \mathbf{b}_i \rangle$ denotes the posterior expectation of the vector containing the regulatory intensities with which transcription factor i influences all the genes in the genome (set to zero for genes that are not bound by that transcription factor).

Algorithm 2 Identifying transcription factors associated with submodules of a network using the regulatory intensities.

Input data: affinity matrix A;
repeat
Compute the eigendecomposition of A, giving eigenvalues λ_i and eigenvectors $E = \{\mathbf{e}_i\}$, $i = 1, \ldots, q$;
Define $B = \{\mathbf{e}_1\}$, $\bar{B} = E - B$
If $\mathbf{e}_i \in \bar{B}$ is such that $|\mathbf{e_j}|^T |\mathbf{e}_i| = 0 \quad \forall \mathbf{e}_j \in B$, include $\mathbf{e_i}$ in B;
until No such $\mathbf{e}_i$ can be found

We use the absolute value of the intensity since for the purpose of identifying submodules we are not interested in the sign of the regulation. According to this formula, then, two transcription factors will have high similarity if they coregulate with high intensity a large number of target genes.

If we assume that there are p independent submodules, with strong internal links, the affinity matrix (5) will be have p blocks on the diagonal (up to a reordering of the rows and columns) showing a very high internal covariance, while the remaining off-diagonal entries will be much smaller. By identifying these blocks, one can then obtain the transcription factors involved in the submodules. The blocks can be obtained by noticing that, for a non-degenerate spectrum (which holds with probability 1), the eigenvectors of C will present a block structure too, so that eigenvectors pertaining to different blocks will have non-zero entries in different positions. By selecting exactly one eigenvector per each block we obtain a set of *clustering eigenvectors*[3], and we can obtain the transcription factors belonging to different modules by considering the nonzero entries of the clustering eigenvectors. Furthermore, the eigenvalues associated with the clustering eigenvectors are monotonically related to the total regulatory intensity associated with the submodule (the sum of all the regulatory intensities of all the links in the network). Therefore, we can use the eigenvalues to rank the various submodules in terms of their importance in the overall network. A strategy to identify the submodules can therefore be obtained as outlined in Algorithm 2.

If the modules are not exactly independent, but links between modules are characterised by low regulatory intensity, we can introduce a sensitivity parameter θ and replace step 3 in algorithm 2 by $|\mathbf{e_j}|^T |\mathbf{e}_i| < \theta \quad \forall \mathbf{e}_j \in B$. As the eigenvectors of a matrix with non-degenerate spectrum are stable under perturbations, we are guaranteed that, for suitably small choices of θ, approximately independent submodules will be found.

In practice, it is often the case in biological networks that there are few submodules of the regulatory network active in a given experimental condition, so that we may expect the submodules identified by the clustering eigenvectors with highest associated eigenvalue to be biologically relevant, while submodules associated with small eigenvalues will be less relevant.

[3] The name is chosen for their analogy with spectral clustering [10].

The simplicity of the algorithm leads to several advantages. For example, by considering the eigenvectors of the dual matrix

$$K_{lp} = \sum_{i=1}^{q} |\langle b_{li} \rangle| \, |\langle b_{pi} \rangle| \,, \tag{6}$$

one can retrieve the genes involved in the submodules.

4 Results

4.1 Data sets

We tested our method on two yeast data sets, the benchmark cell cycle data set of [11] and the recent metabolic cycle data set of [12]. These data sets were analysed in our recent studies [8, 9]. The connectivity data we used in both cases was obtained using ChIP: for the metabolic cycle data, we used the recent ChIP data of [1], while for the cell cycle data we chose to use the older ChIP data of [2] since this combination has been extensively studied in the literature [15, and references therein]. The ChIP data is continuous, but, following the suggestion of [2], we binarised it by giving a one value when the associated p-value was smaller than 10^{-3}. This was shown in [8] to be a reasonable choice of cut-off, as it retained many regulatory relations while keeping the number of false positives reasonable.

4.2 Cell cycle data

Spellman *et al.* [11] used cDNA microarrays to monitor the gene expression levels of 6181 genes during the yeast cell cycle, discovering that over 800 genes are cell cycle-regulated. Cells were synchronised using different experimental techniques. We selected the cdc15 data set, consisting of 24 experimental points in a time sequence.

The connectivity data we used for this data set was that obtained by [2]. In this study, ChIP was performed on 113 transcription factors, monitoring their binding to 6270 genes.

We removed from the data set genes which were not bound by any transcription factor and transcription factors not binding any gene. We also removed the expression data of genes with five or more missing values in the microarray data, leaving a network of 1975 genes and 104 transcription factors.

For the purposes of identifying submodules, we are primarily interested in the regulatory intensities with which transcription factors regulate target genes. Therefore, we will use the model described in Section 2 to obtain posterior estimates for the regulatory intensities b_{nm}. Also, we will be interested primarily in nontrivial submodules, *i.e.* submodules involving more than one regulator.

Identifying submodules using the ChIP data As ChIP monitors only the binding of transcription factors to promoter regions of genes, and not the actual regulation, we may expect that many true positives at the binding level are actually false positives at the regulatory level. For example, the ChIP data of [2], using a p-value of 10^{-3}, gives 3656 bindings involving 104 transcription factors and 1975 genes. However, if we consider the posterior statistics for the regulatory intensities, we see that most of these bindings are not associated with a regulatory intensity significantly different from zero. Specifically, only 1238 bindings are associated with a regulatory intensity greater than twice its posterior error (significant with 95% confidence), and only 749 are significant at 99% confidence level.

This large number of false positives is a serious problem when trying to identify submodules. For example, if we use the naive Algorithm 1 directly on the ChIP data, we obtain only one nontrivial[4] submodule involving 100 transcription factors and 1957 genes. Obviously, the usefulness of such information is very limited.

Identifying submodules using significant regulations Things change dramatically if we construct a binary connectivity matrix by considering only significant regulatory relations. In order to avoid obtaining too large components, we fixed the thresholding parameter to be equal to four. At this stringent significance threshold the network size reduces significantly, as there are now 81 transcription factors regulating a total of 438 genes. More importantly, there are now nine distinct nontrivial submodules of the regulatory network, each involving between two and thirteen transcription factors.

The submodules identified are highly coherent functionally. To appreciate this, we follow [2] and group transcription factors into five broad functional categories according to the function of their target genes. These categories are cell cycle, developmental processes, DNA/RNA biosynthesis, environmental response and metabolism [2, see Figure 5 inset]. We then see that the largest submodule, consisting of 13 transcription factors regulating 117 genes, is largely made up of transcription factors functionally related to the cell cycle. In fact, all of the active transcription factors functionally related to the cell cycle (with the exception of SKN7 and SWI6 which are not involved in any nontrivial module) belong to this submodule. These are ACE2, FKH1, FKH2, MBP1, MCM1, NDD1, SWI4 and SWI5. Among the other transcription factors in the module, three (STE12, DIG1 and PHD1) are associated with developmental processes and the remaining two (RLM1 and RFX1) are associated with environmental response. The presence of these transcription factors in the same module could indicate a coupling between different cellular processes (for example, it is reasonable that cell cycle and cell development could be coupled), but it could also be due to the fact that certain transcription factors may be involved in more than one cellular process,

[4] There are four trivial submodules made up of a single transcription factor regulating genes with only one regulator.

hence rendering the boundaries between functional categories somewhat fuzzy. A graphical representation[5] of this submodule is given in Figure 1.

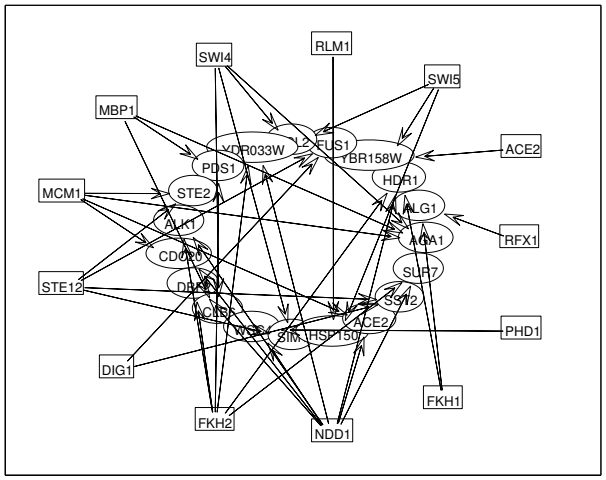

Fig. 1. Graphical representation of the nontrivial part of the cell cycle submodule of the regulatory network obtained by considering only significant regulatory relations. The boxes represent the transcription factors, the inner vertices represent the 19 genes regulated by more than one transcription factor.

The smaller submodules exhibit similar functional coherence. For example, there are four independent submodules involving transcription factors related to cell metabolism, consisting respectively of: ARG80, ARG81 and GCN4; ARO80 and CBF1; LEU3 and RTG3 and DAL82 and MTH1. Other two submodules consist mainly of genes related to environmental response, one including CIN5, MAC1 and YAP6 together with AZF1 (related to metabolism) and the other one including CAD1 and YAP1. The remaining two submodules consist of two transcription factors belonging to different functional categories. The nontrivial part of one of these submodules is shown graphically in Figure 2. As it can be seen, this is a reasonably sized system which could be amenable to a more detailed description.

Identifying submodules using regulatory intensities While considering only significant regulations clearly leads to a significant advantage when trying to

[5] The graphs in this paper were obtained using the MATLAB interface for GraphViz, available at `http://www.cs.ubc.ca/~murphyk/Software/GraphViz/graphviz.html`.

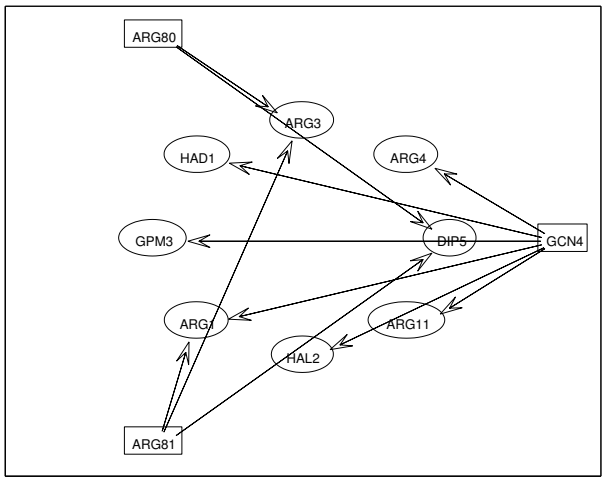

Fig. 2. Graphical representation of one of the submodules of the regulatory network obtained by considering only significant regulatory relations. This submodules is functionally related to the cell metabolism.

identify submodules, a simple thresholding technique as discussed in the previous section clearly does not make use of the wealth of information contained in the regulatory strengths. We therefore studied the cell cycle data using the spectral algorithm described in section 3.2.

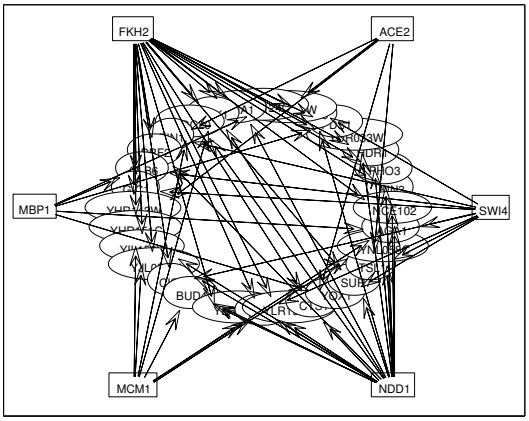

Fig. 3. Graphical representation of the principal submodule obtained by considering the regulatory intensities. All the transcription factors involved in this submodule (indicated in the outer boxes) are key regulators of the cell cycle. The inner vertices represent the genes with more than one significant regulator involved in the submodule.

We constructed the affinity matrix as in (5) by using all regulatory intensities with a signal to noise ratio greater than 2 (95% significance level) and selecting only genes significantly regulated by two or more transcription factors (these are the only ones that will contribute to the off-diagonal part of the covariance). We then applied the submodule finding Algorithm 2 with a sensitivity parameter 0.01. This gave four clustering eigenvectors, yielding submodules involving between seven and two transcription factors each. Ranking these using the eigenvalues associated, we find that the submodules exhibit a remarkable functional coherence. For example, 98.7% of the mass of the first clustering eigenvector is accounted for by six transcription factors. These are ACE2, FKH2, MBP1, MCM1, NDD1 and SWI4 and are all functionally associated with the cell cycle. By considering the genes involved in this submodule, obtained by considering the eigendecomposition of the dual matrix (6), we also recognise some key genes involved in the cell cycle, such as AGA1, CLB2, CTS1, YOX1 and the transcription factor genes ACE2 and SWI5. The nontrivial part of this submodule of the regulatory network is shown in Figure 3. Similarly, the second eigenvector has 99.9% of its mass concentrated on two transcription factors, DAL82 and

MTH1, which are related to carbohydrate/nitrogen metabolism, 99.3% of the third eigenvector's mass is accounted for by AZF1, CUP9 and DAL81, which are related to cell metabolism (CUP9 is also associated with response to oxidative stress), 99% of the mass of the fourth clustering eigenvector is accounted for by LEU3 and STP1, both related to cell metabolism.

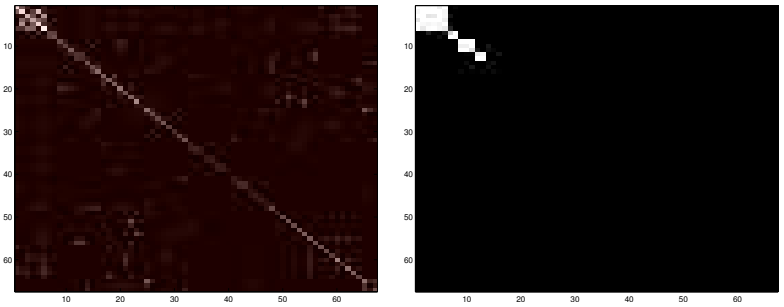

Fig. 4. Graphical representation of the affinity matrix obtained using the regulatory intensities for the cell cycle data set(*left*) and block structure obtained from the submodules found using the spectral Algorithm 2. One strongly interconnected submodule is evident in the top left corner of the affinity matrix; the other submodules are associated with much weaker interactions and are hard to appreciate at a glance.

A major difference with the naive submodule finding Algorithm 1 is the non-exhaustive nature of the spectral algorithm. Specifically, while the naive algorithm will assign each transcription factor represented in the network to exactly one (possibly trivial) submodule, most transcription factors are not included into any submodule by the spectral algorithm. This can be understood by considering the structure of the affinity matrix, which is shown graphically in Figure 4, *left*. While there is one evident block with very high internal covariance in the top left corner (representing the dominant clustering eigenvector associated with the cell cycle), the other submodules are not easily appreciated, since they are associated with much weaker regulatory intensities. The block structure given by the submodules is shown graphically in Figure 4 *right*. Notice however that most transcription factors are not associated with any submodule, indicating that they do not appear to be key in any cellular process going on during the cell cycle.

4.3 Metabolic cycle data

Tu *et al.* used oligonucleotide microarrays to measure gene expression levels during the yeast metabolic cycle, *i.e.* glycolitic and respiratory oscillations following

a brief period of starvation. The samples were prepared approximately every 25 minutes and covered three full cycles, giving a total of 36 time points [12].

The connectivity we used to analyse this data set was obtained integrating the two ChIP experiments of Lee *et al.* [2] and Harbison *et al.* [1], resulting in a very large network of 3178 genes and 177 transcription factors. By integrating the two datasets, we capture the largest number of potential regulatory relations, which also implies we are introducing a large number of false positives. It is not surprising then that trying to identify submodules directly from the ChIP data leads to a single huge module including all transcription factors and all genes.

Perhaps more surprisingly, the situation does not improve much if we consider only regulations with a high significance level (signal to noise ratio greater than four). Although the number of significant regulations is much smaller than the number of potential regulations (1826 versus 7082), the resulting network still appears to be highly interconnected, so that the application of the naive algorithm again yields one very large submodule (134 transcription factors) and two small submodules containing two transcription factors each. These ones are CST6 and SFP1, two transcription factors which may be loosely related to metabolism (CST6 regulates genes that utilise non optimal carbon sources, while SFP1 activates ribosome biogenesis genes in response to various nutrients) and A1(MATA1) and UGA3, which do not appear to have an obvious functional relationship.

We get a completely different picture if we use the information contained in the regulatory strengths. If we again construct an affinity matrix by retaining the regulatory strengths of all regulations which are significant at 95% for genes regulated by at least two transcription factors, the spectral submodule finding Algorithm 2 (again with sensitivity parameter set to 0.01) returns seven non-trivial submodules.

Somewhat surprisingly, the first clustering eigenvector is again related to the cell cycle: 96.6% of its mass is concentrated on the ten transcription factors ACE2, FKH2, MBP1, MCM1, NDD1, SKN7, STB1, SWI4, SWI5 and SWI6, which are all well known key players of the yeast cell cycle. This seems to add support to the hypothesis, advanced by Tu *et al.*, that the metabolic cycle and the cell cycle might be coupled [12]. The functional coherence of the other submodules is less clear: while GTS1 and RIM101, which account for 99.8% of the mass of the second clustering eigenvector, are both involved in sporulation, the functional annotations of the transcription factors involved in other submodules are less coherent. For example, the coupling between MSS11 (which regulates starch degradation) and WAR1 (which promotes acid and ammonia transporters) is plausible but may need further experimental validation before being accepted. A graphical representation of the submodule formed by GTS1 and RIM101 is given in Figure 5.

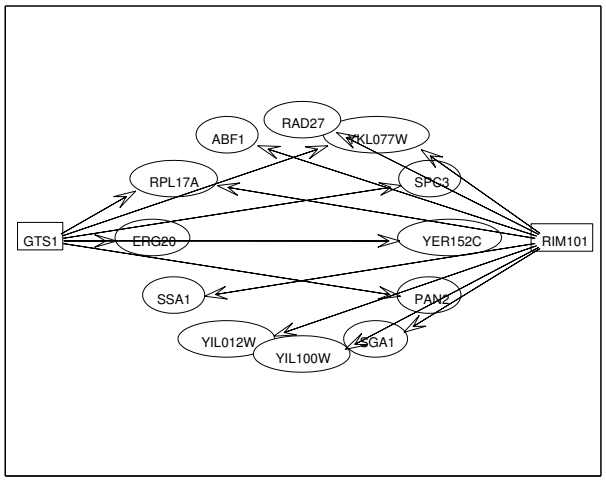

Fig. 5. Graphical representation of the submodule of the metabolic cycle given by GTS1 and RIM101, two transcription factors involved in regulating sporulation.

5 Discussion

In this paper we proposed two algorithms to identify approximately independent submodules of the cellular regulatory network. Both methods rely on having genome-wide information on the intensity with which transcription factors regulate their target genes, obtained for example by using the recent model proposed in [9]. While the first algorithm is a simple exhaustive search, the second is more subtle, being based on the spectral decomposition of an affinity matrix between transcription factors, and is somewhat related to the algorithm proposed in [10] for the automatic detection of non-convex clusters.

Experimental results obtained using the algorithms on two yeast data sets reveals that both methods can find biologically plausible submodules of the regulatory network, and in many cases these submodules are of small enough size to be amenable to be modelled in a more detailed fashion. The two algorithms have complementary strengths: while the naive search algorithm has the advantage of assigning each transcription factor to a unique submodule, many transcription factors are not assigned to any module by the spectral algorithm. On the other hand, the functional coherence of the submodules identified by the spectral algorithm seems to be higher in the examples studied, and sensible submodules are found even when the network is too interconnected for the naive search to yield any submodules.

Another popular method to cluster graphs which has been extensively applied to biological problems is the Markov Cluster Algorithm (MCL), which was used successfully to find families of proteins from sequence data [16]. However, this algorithm is designed for undirected graphs with an associated similarity matrix, while the graphs obtained from regulatory networks are naturally directed (with arrows going from transcription factors to genes). Even if we marginalise the genes by considering an affinity matrix between transcription factors, this is generally not a consistent similarity matrix, making the application of MCL very hard. Bearing in mind the largely exploratory nature of finding submodules of the regulatory network, we preferred to use simpler and more interpretable methods.

Acknowledgements

The authors gratefully acknowledge support from a BBSRC award "Improved processing of microarray data with probabilistic models".

References

1. C. T. Harbison et al., Nature **431**, 99 (2004).
2. T. I. Lee et al., Science **298**, 799 (2002).
3. T. Schlitt and A. Brazma, FEBS letts **579**, 1859 (2005).
4. N. M. Luscombe et al., Nature **431**, 308 (2004).
5. N. A. Monk, Biochemical Society Transactions **31**, 1457 (2003).
6. X. Xie et al., Nature **434**, 338 (2005).
7. R. Martone et al., Proceedings of the National Academy of Sciences USA **100**, 12247 (2003).
8. G. Sanguinetti, M. Rattray, and N. D. Lawrence, A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription, To appear in *Bioinformatics*, 2006.
9. G. Sanguinetti, N. D. Lawrence, and M. Rattray, Probabilistic inference of transcription factors concentrations and gene-specific regulatory activities, Technical Report CS-06-06, University of Sheffield, 2006.
10. G. Sanguinetti, J. Laidler, and N. D. Lawrence, Automatic determination of the number of clusters using spectral algorithms, in *Proceedings of MLSP 2005*, pages 55–60, 2005.
11. P. T. Spellman et al., Molecular Biology of the Cell **9**, 3273 (1998).
12. B. P. Tu, A. Kudlicki, M. Rowicka, and S. L.McKnight, Science **310**, 1152 (2005).
13. J. C. Liao et al., Proceedings of the National Academy of Sciences USA **100**, 15522 (2003).
14. O. Alter, P. O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. USA **97**, 10101 (2000).
15. A.-L. Boulesteix and K. Strimmer, Theor. Biol. Med.Model. **2**, 1471 (2005).
16. A.J.Enright, S. van Dongen, and C. Ouzounis, Nucleic Acids Research **30**, 1575 (2002).