



PUBLIC

SIXTH FRAMEWORK PROGRAMME

Specific Targeted Research Projects

PRIORITY 1

LIFE SCIENCES, GENOMICS AND BIOTECHNOLOGY FOR HEALTH

Contract no: LSHG-CT-2004-512143

DIAMONDS

Dedicated Integration And Modelling Of Novel Data and prior knowledge to enable Systems biology

EU Deliverable

D3.3

Algorithms for the analysis of periodicity in correlated gene expression data sets

Due Date: 1 July 2006

Delivery Date: 16 August 2006

Version 2

Partner responsible: Søren Brunak (**CBS**)



Diamonds D3.3

Algorithms for the analysis of periodicity in correlated gene expression data sets

CONTRIBUTIONS:

Søren Brunak, Thomas S. Jensen & Ulrik de Lichtenberg
Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of
Denmark, Building 208, DK-2800 Lyngby, Denmark

1 Executive Summary

Cell cycle-regulated transcription is one of the focus areas of the DIAMONDS project and this WP3 deliverable reports on the development and deployment of state-of-the art computational methods to the cell cycle microarray expression data collected and generated in WP1.

2 Background

It has been clear for many years that certain genes are expressed only at specific stages of the cell cycle (e.g. the cyclins and the histones). These genes consequently exhibit a periodic pattern of expression when monitored during consecutive cell cycles. The extent of this transcriptional regulation was, however, unclear until the publication of two genome-wide DNA microarray studies of the *Saccharomyces cerevisiae* cell cycle (Cho et al., 1998; Spellman et al., 1998). These studies concluded that 400-800 genes fluctuated in expression during the cell cycle, but did not agree on exactly which genes to consider as periodically expressed. Similar investigations were later performed using microarrays in human fibroblasts (Cho et al., 2001), in HeLa cells (Whitfield et al., 2002), in fission yeast (Rustici et al., 2004; Peng et al., 2005; Oliva et al., 2005) and in the plant *Arabidopsis thaliana* (Menges et al., 2003). Each of these studies have aimed at defining the cell cycle regulated (or periodically expressed) subset of the genome in each organism.

From the beginning it was apparent that different experiments, when analyzed alone, yielded different results and this has in turn inspired the development of a large number of computational methods for identifying the genes that fluctuate most significantly with progression through the cycle (Spellman et al., 1998; Zhao et al., 2001; Langmead et al., 2002, 2003; Johansson et al., 2003; Lu et al., 2004; Luan and Li, 2004; Wichert et al., 2004; de Lichtenberg et al., 2005a; Ahdesmäki et al., 2005; Chen, 2005; Willbrand et al., 2005; Qiu et al., 2006; Ahnert et al., 2006; Andersson et al., 2006). Most of these have been applied to the budding yeast data (Cho et al., 1998; Spellman et al., 1998) and the lack of overlap between the predictions from different methods have led to a widespread confusion over the number and identity of cell cycle regulated (or periodically expressed) genes in budding yeast. Recently, three different experimental studies of the fission yeast cell cycle have also arrived at different and only partially overlapping subsets of the genome.

In the context of the DIAMONDS project, we have been working hard to find the best way of analyzing the cell cycle gene expression data and to resolve the apparently conflicting evidence from different experimental and computational studies. We report here on this work.

3 Results

3.1 Benchmarking computational methods

Virtually all computational methods for identifying periodically expressed genes from microarray data have been applied to the three original *S. cerevisiae* time-series published by Spellman et al. (1998). Unfortunately, the predictions from these methods are far from overlapping and although most studies agree that there is in the order of 300-800 periodically expressed genes in budding yeast, they do not agree on their identity. In fact, more than 1800 genes have been proposed in total, equivalent to every third gene in the genome.

The key problem is that no external benchmark set (a set of genes known or expected to be cell cycle regulated) existed to settle which methods work best. In order to resolve the issue, we (DIAMONDS partner 5, Søren Brunak) therefore collected sets of genes for which there was some independent source of evidence which would suggest cell cycle-regulated transcription (de Lichtenberg et al., 2005a) and measured the ability of each method to identify genes from those sets. A good method was thus defined as one that identifies genes for which other independent experimental evidence suggest cell cycle regulation.

The results were highly surprising and showed that the majority of the methods developed after the original approach by Spellman et al. (1998) in fact performed much worse (de Lichtenberg et al., 2005a).

For use in DIAMONDS, we developed our own algorithm and showed that it performs as good or better than all other methods (de Lichtenberg et al., 2005a). The method consists of two statistical test which separately quantifies the significance of periodicity and regulation for each gene. This aspect of our method also allowed us to explain the poor performance of many of the newly developed methods. It turns out that the performance of a method is strongly related to whether it takes into account the magnitude or significance of regulation of the gene. The magnitude of regulation is part of the methods by Spellman et al. (1998), Johansson et al. (2003) and de Lichtenberg et al. (2005a), which top the ranking, whereas the later methods only assess the periodicity and consequently loose part of the signal. The results of the benchmark analysis was confirmed when looking at the overlap between genes identified as periodic in individual experiment: the magnitude-dependent methods yield a better overlap than do methods which only take into account the shape of the profile (de Lichtenberg et al., 2005a).

Since the publication of the benchmark analysis, several new methods have appeared in the literature (Ahdesmäki et al., 2005; Chen, 2005; Willbrand et al., 2005; Qiu et al., 2006; Ahnert et al., 2006; Andersson et al., 2006). Figure 1 shows an updated version of the performance of all methods published to date on benchmark set B2, which consists of 352 genes whose promoters are associated with at least one known cell cycle transcription factor (de Lichtenberg et al., 2005a; Simon et al., 2001; Lee et al., 2002).

The results show that none of the newly published methods yield any improvement over the original analysis and confirm that our method represents state-of-the-art. In collaboration with DIAMONDS partner 11 (Jürg Bähler), we have recently performed a similar benchmark for ten experiments on the fission yeast cell cycle which show that in our method in all cases yield results that are as good or better than the original analysis of the data.

In conclusion, the benchmark shows that we have developed a computational method for identifying cell cycle-regulated genes, which ranks among the best. The algorithm has been implemented into the super-computing facility run by DIAMONDS partner 5 (Søren Brunak) and will be applied to the data generated and collected in WP1. Whereas the results of the method will all be made available to both DIAMONDS

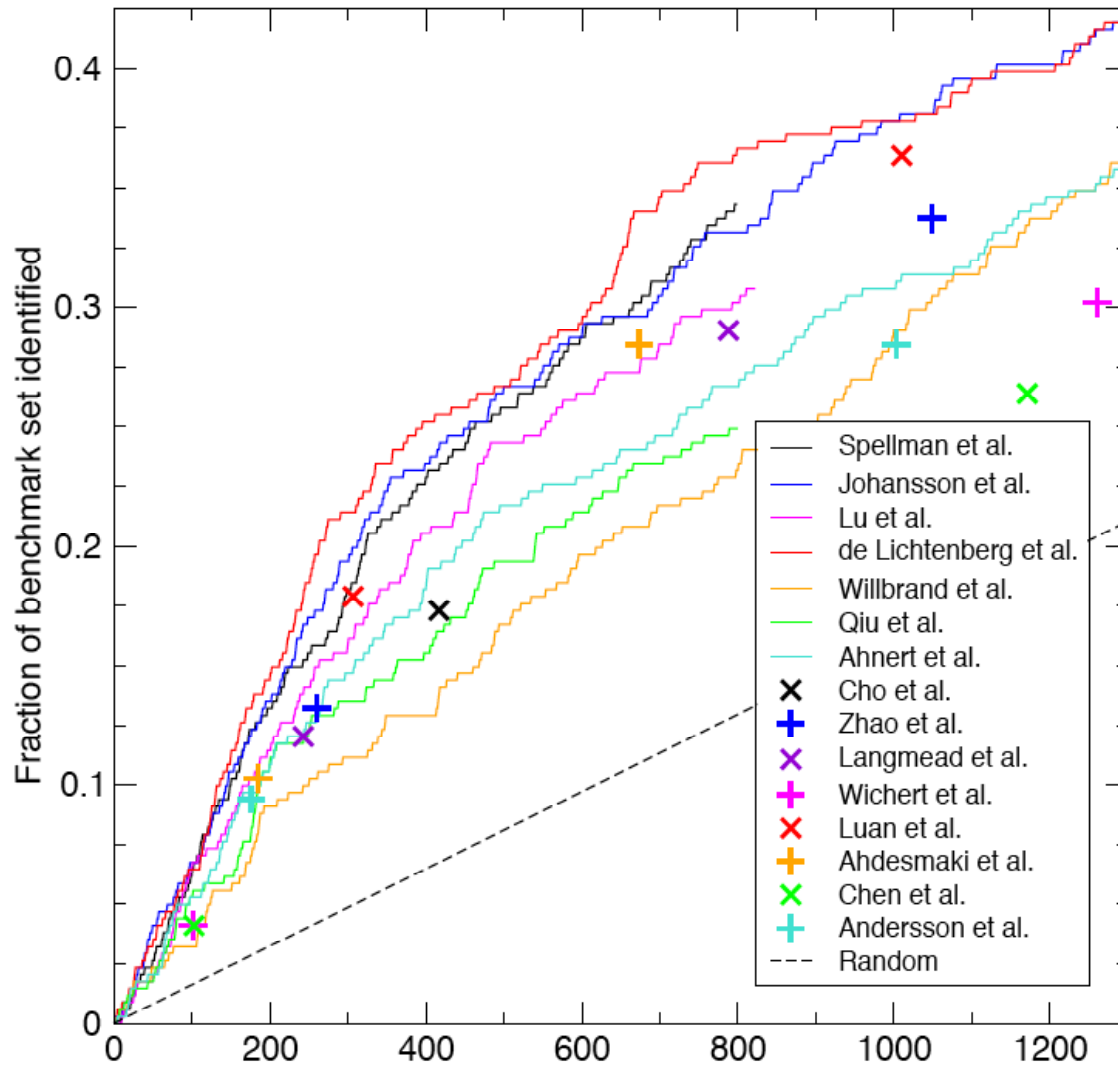


Figure 1: **Comparison of published methods.** The fraction of the benchmark set that is identified is plotted as a function of gene rank for each method, experiment, and benchmark set. Methods which published a ranked list of genes appear as a curve, whereas unranked gene lists appear as points. Random performance is shown as a black dotted line.

partners and the general community, the algorithm itself is not suited for incorporation into routine applications as it relies on permutations that are computationally heavy. For this, we instead recommend a fourier-scoring approach similar to that used by Spellman et al. (1998), as it performs nearly as good and is computationally much easier to handle.

3.2 Application of the new method to experimental data

The algorithms developed for this DIAMONDS deliverable (de Lichtenberg et al., 2005a) has been(or will be) applied to the many gene expression data sets generated or

collected in WP1. For each gene, the method quantifies the significance of regulation and periodicity of its expression profile as well as identifies the time of peak expression. The method can analyze single experiments, but also combine the evidence for any set of time-series. Based on the combined sum of experimental evidence within an organism, the goal is to transform the raw experimental data from WP1 into information of the periodicity and temporal behavior of each gene in the genome.

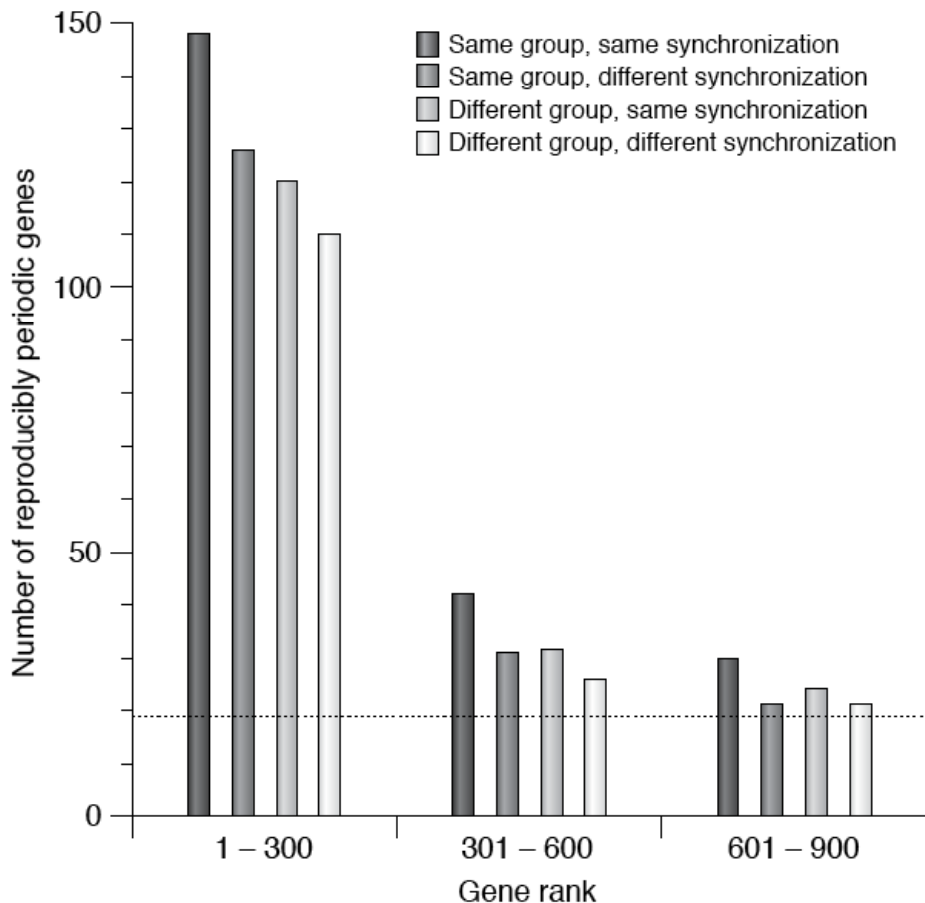


Figure 2: Reproducibility in single microarray experiments. Reproducibility of genes identified in two experiments analysed by the method of de Lichtenberg et al.(2005a). Each bar shows the average number of overlapping genes among two different experiments analyzed individually when using the 300 Highest ranking genes from each experiment (left), or using the genes ranked from 301-600 (middle) and 601-900 (right). The comparisons are subdivided based on whether the experiments were performed in the same laboratory and by using the same protocol for cell-cycle synchronization. There is good reproducibility among the 300 highest ranking genes, but the reproducibility drops close to random expectation (19 genes) for genes in the second and third sets.

In budding yeast, we have analyzed data generated in our own lab (de Lichtenberg et al., 2005b), data from the three original experiments (Cho et al., 1998; Spellman

etal.,1998), as well as data from Linda Breeden's group at the Fred Hutchinson Cancer Research Center(unpublished results). In collaboration with DIAMONDS partner 11 (Jürg Bähler), we have also performed a comprehensive analysis of all ten experiments available on the fission yeast cell cycle (Marguerat et al., 2006). These include the five experiments performed by partner 11 (Rustici et al., 2004), as well as two smaller data sets performed by other groups (Peng et al., 2005; Oliva et al., 2005). We are currently working on similar analysis of publicly available data sets on the human and plant cell cycle.

In all cases, our results have been as good or better than previous analyses and the combined set of analysis offers a unique platform for future work in DIAMONDS. An additional advantage is that the data is consistent across organisms, because it has all been generated with the same state-of-the-art method.

A number of conclusions can be drawn from these analyses: single microarray experiments are subject to a considerable level of noise and many independent experiments are needed to extract the underlying signal. As part of the collaboration with partner 11 (Jürg Bähler), we recorded the overlap between the 300 highest ranking genes in all pairwise comparisons of the ten experiments and found that the typical level of reproducibility is less fifty percent (Figure 2).

Interestingly, however, the biases from experiments performed in different labs or via different synchronization techniques were relatively modest and of similar magnitude. Taken together, these results indicate that the lack of overlap between different studies arises mainly because most groups analyze only their own (small) data set and because each group uses their own computational method (i.e. definition of periodicity). Analyzing all data in combination yields the most reliable results and draws a consistent picture of the transcriptional dynamics in each organism.

4 Conclusions

A computational algorithm has been developed for the identification of cell cycle regulated genes and a thorough benchmark shows that this algorithm represents state-of-the-art. The method has been applied to the data generated or collected in WP1 and both the algorithm as well as the results will be made available to both DIAMONDS partners and the scientific community. In conclusion, the requirements for this deliverable are fulfilled.

5 Perspectives

In connection with the generation and collection of experimental data sets in WP1, this and other algorithms related to DIAMONDS will create a unique source of high quality data on the regulation of the cell cycle across different organism. Apart from use in the context of DIAMONDS, we envision that these data will benefit the entire cell cycle

community. Already now, results generated with our computational method de Lichtenberg et al. (2005a) have been used in work outside DIAMONDS (Gavin et al., 2006; Sopko et al., 2006).

Publications and resources related to this deliverable

- **Comparison of computational methods for the identification of cell cycle regulated genes, de Lichtenberg, U, Jensen, LJ, Fausboll, A, Jensen, TS, Bork, P and Brunak, S, Bioinformatics, 2005, 21(7):1164-1171.**
This paper describes the method developed by partner 5 and presents the first benchmark of computational methods for identifying cell cycle regulated genes.
- **New weakly expressed cell cycle-regulated genes in yeast, de Lichtenberg, U, Wernersson, R, Jensen, TS, Nielsen, HB, Fausbøll, A, Schmidt, P, Hansen, FB, Knudsen, S, and Brunak, S, Yeast, 2005, 22(15):1191-1201.**
This paper describes the data generated by partner 5 to which a modified version of the permutation-based method was applied to identify weakly expressed cell cycle regulated genes.
- **The more the merrier: Comparison and integrative analysis of microarray studies on cell-cycle-regulated genes in schizosaccharomyces pombe, Marguerat, S, Jensen, TS, de Lichtenberg, U, Wilhelm, BT, Jensen, LJ and Bähler, J, Yeast, 2006, 3(4):261-77.**
This paper describes the combined analysis of all ten experiments in fission yeast performed in collaboration between partners 5 and 11.
- **www.cbs.dtu.dk/cellcycle.**
This website contains the final results of our analyses in addition to benchmark sets and supplementary information on our work.

References

- Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6:117.
- Ahnert, S. E., Willbrand, K., Brown, F. C. S., and Fink, T. M. A. (2006). Unbiased pattern detection in microarray data series. *Bioinformatics*, 22(12):1471–1476.
- Andersson, C. R., Isaksson, A., and Gustafsson, M. G. (2006). Bayesian detection of periodic mRNA time profiles without use of training examples. *BMC Bioinformatics*, 7:63.
- Chen, J. (2005). Identification of significant periodic genes in microarray gene expression data. *BMC Bioinformatics*, 6:286.
- Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73.
- Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, 27(1):48–54.
- de Lichtenberg, U., Jensen, L. J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005a). Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics*, 21(7):1164–1171. doi:10.1093/bioinformatics/bti093.
- de Lichtenberg, U., Wernersson, R., Jensen, T. S., Nielsen, H. B., Fausbøll, A., Schmidt, P., Hansen, F. B., Knudsen, S., and Brunak, S. (2005b). New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, 22(15):1191–1201.
- Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and SupertiFurga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636.
- Johansson, D., Lindgren, P., and Berglund, A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, 19(4):467–473.
- Langmead, C. J., McClung, C. R., and Donald, B. R. (2002). A maximum entropy algorithm for rhythmic analysis of genome-wide expression patterns. *Proc IEEE Comput Soc Bioinform Conf*, 1:237–245.

Langmead, C. J., Yan, A. K., McClung, C. R., and Donald, B. R. (2003). Phase-independent rhythmic analysis of genome-wide expression patterns. *J Comput Biol*, 10(3-4):521–536.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804.

Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. (2004). Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res.*, 32:447–455.

Luan, Y. and Li, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3):332–339.

Marguerat, S., Jensen, T. S., de Lichtenberg, U., Wilhelm, B. T., Jensen, L. J., and Bähler, J. (2006). The more the merrier: Comparison and integrative analysis of microarray studies on cell-cycle-regulated genes in *Schizosaccharomyces pombe*. *Yeast*. 2006 Mar;23(4):261-77.

Menges, M., Hennig, L., Grussem, W., and Murray, J. A. H. (2003). Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Mol Biol*, 53(4):423–42.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Fitcher, B., and Leatherwood, J. (2005). The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol*, 3(7):e225.

Peng, X., Karuturi, R. K. M., Miller, L. D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L.-S., Liu, E. T., Balasubramanian, M. K., and Liu, J. (2005). Identification of cell cycle-regulated genes in fission yeast. *Mol Biol Cell*, 16(3):1026–42.

Qiu, P., Wang, Z. J., and Liu, K. J. R. (2006). Polynomial model approach for resynchronization analysis of cell-cycle gene expression data. *Bioinformatics*, 2(8):959–966.

Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, 36(8):809–817.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708.

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S. G., Cyert, M., Hughes, T. R., Boone, C., and Andrews, B. (2006). Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell*, 21(3):319–330.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, 13:1977–2000.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

Willbrand, K., Radvanyi, F., Nadal, J.-P., Thiery, J.-P., and Fink, T. M. A. (2005). Identifying genes from up-down properties of microarray expression series. *Bioinformatics*, 21(20):3859–3864.

Zhao, L. P., Prentice, R., and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5631–5636.