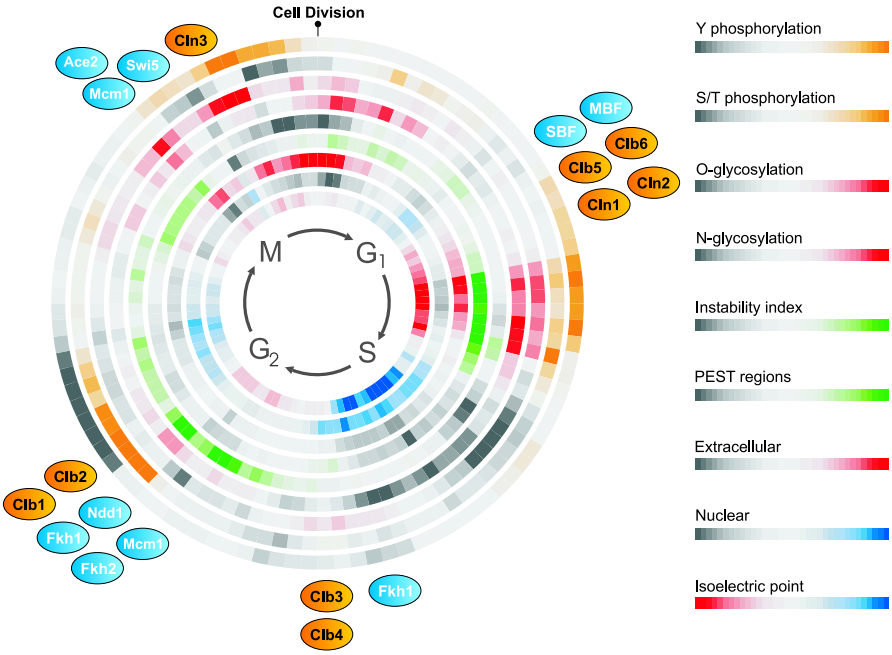
Supplementary Information for

Protein Feature Based Identification of Cell Cycle Regulated Proteins in Yeast

Ulrik de Lichtenberg, Thomas S. Jensen, Lars J. Jensen and Søren Brunak

Journal of Molecular Biology (2003) **329**, 663–674



Summary

DNA microarrays have been used extensively to identify cell cycle regulated genes in yeast, however, the overlap in the genes identified is surprisingly small. We show that certain protein features can be used to distinguish cell cycle regulated genes from other genes with high confidence (features include protein phosphorylation, glycosylation, subcellular location and instability/degradation). We demonstrate that co-expressed, periodic genes encode proteins which share combinations of features, and provide an overview of the proteome dynamics during the cycle. A large set of novel putative cell cycle regulated proteins were identified, many of which presently have no known function.

Training set

Training of neural networks for classification requires a set of examples with representatives of each category. In this study only two categories were used, namely "cell cycle regulated protein" and "non cell cycle regulated protein". Selection of training examples was based on a periodicity analysis of the publicly available DNA microarray data sets compiled by Spellman *et al.*[2, 1], to identify periodicly as well as non periodicly expressed genes/proteins.

A Fourier scoring system inspired by Spellman *et al.*[2] was used, where each gene i is assigned a score D_i based on its temporal expression profile during the experiment, with cell cycle frequency $\omega = \frac{2\pi}{T}$:

$$D_i = \sqrt{(\sum_{\mathrm{t}} sin(\omega t) x_i(t))^2 + (\sum_{\mathrm{t}} cos(\omega t) x_i(t))^2}$$

The cell cycle periods, T, estimated by Zhao *et al.*[3] were used (58 min for the α-factor experiment, 115 min for the $Cdc15$ experiment and 85 min for the $Cdc28$ experiment), and a combined Fourier score, F_i, was computed as:

$$F_i = \frac{(D_{i,\alpha} + 0.8 \cdot D_{i,cdc15} + D_{i,cdc28})}{3}$$

The contribution from the $Cdc15$ experiment was scaled in the combined score, because this experiment covers 2.5 cell cycles, whereas the α-factor and $Cdc28$ experiments cover only two (using the Zhao *et al.*[3] estimates).

Figure 1A shows the genome-wide distribution of combined Fourier scores. From this we selected the lowest scoring 556 genes (thresholding at 0.75) to use as our set of "non cell cycle regulated proteins", whose genes display no periodic regulation during the cell cycle.

The discrepancies between DNA microarray studies discussed in the paper underlines the difficulties in selecting a high confidence data set of periodicly expressed proteins. To identify a conservative threshold, we used the 104 known periodic genes listed by Spellman *et al.*[2] to estimate the overall number of periodic transcripts:

$$N_{estimated} = \frac{N_{included}}{M_{included}/M_{total}}$$

The estimated number of periodic transcripts, $N_{estimated}$, was based on the number of genes that score above a certain threshold, $N_{included}$, divided by the fraction of the 104 known genes included above a that threshold, $M_{included}/M_{total}$. Assuming that the periodicity score distribution of the known genes is representative of the entire group of periodic genes, this estimate would be expected to remain constant, so long as the threshold is set high enough to exclude false positives[4].

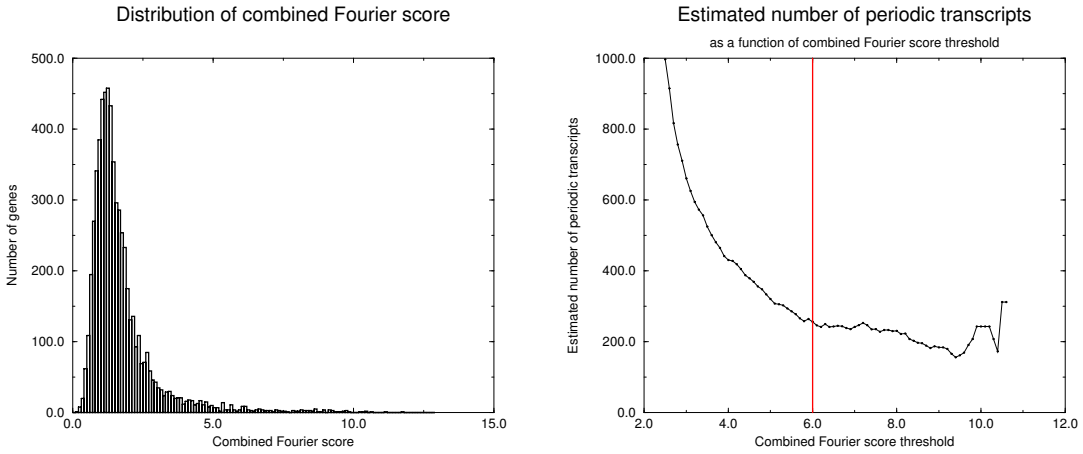


Figure 1: ***A)*** *Genome-wide distribution of Fourier scores.* ***B)*** *Estimated number of periodic transcripts as a function of combined Fourier score threshold. The red line indicates our cutoff. At high Fourier scores large fluctuations are seen, probably due to the poor statistic foundation (very few genes are included).*

The estimate is plotted in Figure 1B and displays a plateau at high thresholds, with an exponential rise towards lower values. Based on this, we applied our most conservative threshold at the end of the plateau (at 6.0) including 115 significantly periodic genes, which most likely include no false positives. To ensure not only periodicity, but also consistent behavior over multiple cycle, we required the Pearson correlation coefficient between the expression profiles of the first and the second cycle to be above 0.4, thereby excluding 18 genes. This procedure resulted in a high confidence set of 97 cell cycle proteins, encoded by genes with strongly periodic and self-consistent expression profiles over three DNA microarray experiments.

The sets of "periodic" and "non-periodic" genes/proteins are available for download from the website, www.cbs.dtu.dk/cellcycle

Neural Network Training

Protein features were derived for each of the proteins in our data set (see above) resulting in a set of 645 examples of "cell cycle regulated proteins" and "non cell cycle regulated proteins" (8 proteins were discarded due to incomplete feature predictions). The protein sequences (translated ORFs) corresponding to the entire *S. cerevisiae* genome were downloaded from the *Saccharomyces Genome Database* (SGD), http://genome-www.stanford.edu/Saccharomyces/. Table 1 summarizes the features explored in this project.

Feature	Tool/program	Reference
Ser/Thr Phosphorylation	NetPhos	(Blom et al.[5]) www.cbs.dtu.dk/services/NetPhos/
Tyr Phosphorylation	NetPhos	(Blom et al.[5]) www.cbs.dtu.dk/services/NetPhos/
PEST sequences	PESTfind	(Reichsteiner and Rogers[6]) www.at.embnet.org/embnet/tools/bio/PESTfind/
Signal Peptides	SignalP	(Nielsen et al.[7]) www.cbs.dtu.dk/services/SignalP/
N-linked Glycosylation	NetNGlyc	(Gupta et al., manuscript in preparation) www.cbs.dtu.dk/services/NetNGlyc/
O-GlcNAc Glycosylation	YinOYang	(Gupta et al., manuscript in preparation) www.cbs.dtu.dk/services/YinOYang/
O-GalNAc Glycosylation	NetOGlyc	(Hansen et al.[8]) www.cbs.dtu.dk/services/NetOGlyc/
Transmembrane helices	TMHMM	(Krogh et al.[9]) www.cbs.dtu.dk/services/TMHMM/
Subcellular Localization	PSORT	(Nakai and Horton[10]) psort.nibb.ac.jp/
Isoelectric Point	ProtParam	www.expasy.ch/tools/protparam.html
Instability Index	ProtParam	www.expasy.ch/tools/protparam.html
Extinction Coefficient	ProtParam	www.expasy.ch/tools/protparam.html
GRAVY	ProtParam	www.expasy.ch/tools/protparam.html
Aliphatic Index	ProtParam	www.expasy.ch/tools/protparam.html
Amino Acid Composition	ProtParam	www.expasy.ch/tools/protparam.html
Protein Sequence Length	ProtParam	www.expasy.ch/tools/protparam.html
Number of pos. residues	ProtParam	www.expasy.ch/tools/protparam.html
Number of neg. residues	ProtParam	www.expasy.ch/tools/protparam.html

Table 1: *Protein features explored individually and in combination in this work.*

Isoelectric point, instability index, extinction coefficient, GRAVY and aliphatic index were all represented by a single value for the entire protein. Subcellular localization contains eleven categories with a probability for each, i.e. an eleven dimensional input vector for each protein. Similarly, the amino acid composition contained 21 input values — one for each amino acid. Other features are residue specific predictions, where the algorithm outputs a prediction for each amino acid in the protein. We summed the predictions and devided by the number of amino acids in the sequence, to avoid the length dependence. This was done for phosphorylation, glycosylation, PEST sequences, signal peptides and transmembrane helixes. Predicted and calculated features were obtained for all translated ORFs in the genome, as desribed above. For each feature we substracted the genome-wide mean and divided by the standard deviation to bring all input data in the same numerical range.

Three-fold cross validation was used (see Figure 2) to divide the data set in three different ways, each with 430 sequences for training and 215 for independent evaluation of the classification performance. Consequently, three networks were trained for each input combination and the classification performance was evaluated as the average Matthews correlation coefficient[11] of the three test sets (in such a way that subset B was only used to test the network trained on subsets A+C, subset A only used to test the one trained on subset B+C, etc., according to Figure 2).

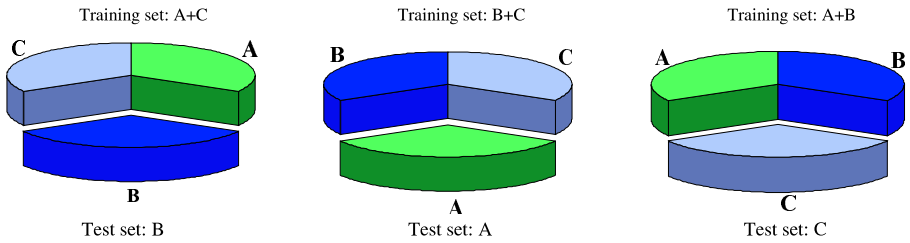


Figure 2: *Schematic illustration of the three-fold cross validation principle.*

The training was performed in an iterative fashion (similar to that of Jensen *et al.*[12]), selecting for the most discriminative features. The features that proved most discriminative in combinations of two were used to construct new combinations of three features, from which the best were selected to form combinations of four, etc. stopping at combinations of six input features (note that some features were encoded as multiple inputs, such as the 11 categories of the subcellular location predictor PSORT). The iteration was continued with the following procedure, until no improvements were possible:

- Optimize the number of hidden neurons
- Test all input combinations obtainable by adding or removing a feature
- Pick the best new input combinations
- Repeat

This iterative selection approach resulted in a number of input combinations and network architectures from which the five best were selected. In the paper we have reported four unique input combinations, because two of the five use identical features, but different network architectures. Each input combination represents three independently trained and tested neural networks (three-fold cross validation) and all 15 networks were combined into a neural network ensemble for improved performance. To put equal weight on all networks, the distributions of test set scores were ranked and used as conversion tables for output from individual networks. We then simply computed the average of the rank converted output scores from all 15 networks. Hence all networks contributed equally to the final scores.

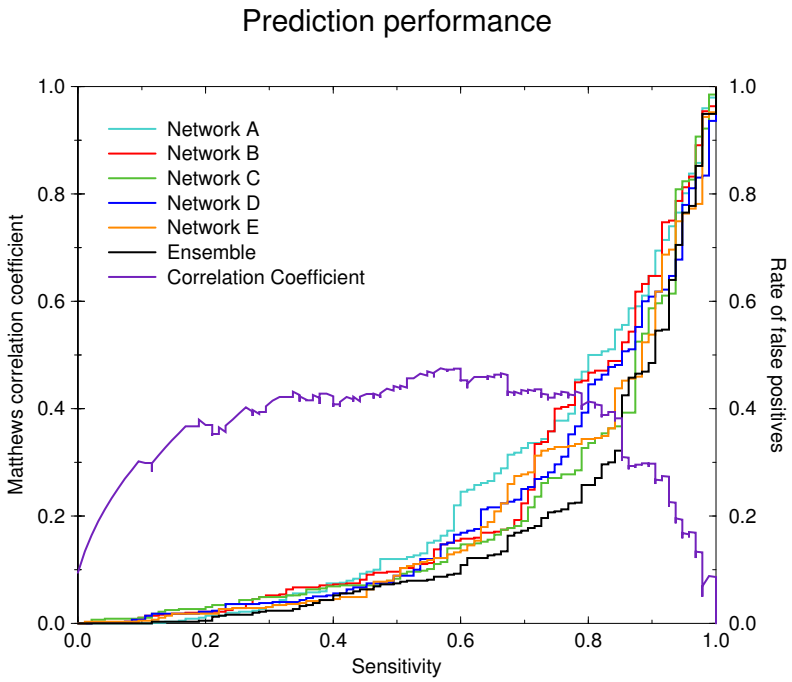


Figure 3: *ROC-curve of the performance of each network and the ensemble, plotted along with the Matthews correlation coefficient*[11]*. The ensemble outperforms all the individual networks.*

Figure 3 contains six ROC-curves, each showing the sensitivity versus the rate of false positives (See paper for definition of sensitivity). From the three independent tests a curve was constructed for each individual input combination (four different combinations, two with different network architecture). Similarly, a curve was constructed for the entire ensemble. It should be noted that this was done such that no network was tested on sequences also used to train that network. As can be seen in Figure 3, the ensemble outperforms all of the individual input neural networks. This is a well known phenomenon with neural networks, that juries, ensembles or averaging of many independently trained networks improve the performance. Figure 3 also contains the Matthews correlation coefficient as a function of the sensitivity. An important conclusion from this plot is that in the regions of low false positive rate, the method also has a low sensitivity. The method should therefore be expected to miss a considerable number of cell cycle regulated proteins. The output should thus be used to support other evidence or to guide new experiments.

The ensemble of trained neural networks was used to predict cell cycle regulated proteins in the entire *S. cerevisiae* proteome (set of all translated ORFs). However, as stated in the paper, proteins contained in the training set or considered "spurious" or "very hypothetical" by Wood *et al.*[13] were removed from the predictions. Each protein was assigned a score between 0 and 1, where high scores are indicative of a cell cycle role for the protein, whereas low scores are less conclusive.

Identification of weakly expressed genes

We examined the intensity distributions of different sets of proposed cell cycle regulated genes by computing the median fluorescence intensity of each gene in each of the three experiments (α-factor, $Cdc28$ and $Cdc15$). The data of Cho *et al.*[1] was used directly, whereas the Spellman *et al.*[2] raw data was normalized with the nonlinear Q-spline method developed by Workman *et al.*[14]. Median intensities were converted into rank statistics for every experiment and the median rank over all three experiments was used as measure of *median intensity*.
Figure 4 shows distributions of the estimated median intensities for the periodic genes identified in the three microarray studies, genes suggested by our ensemble and the entire *S. cerevisiae* genome. From all of these sets we removed genes annotated by Wood *et al.* as "spurious" or "very hypothetical". The distributions were thus based on 400 genes from Cho *et al.*[1], 734 genes from Spellman *et al.*[2], 246 genes from Zhao *et al.*[3], the 500 highest scoring genes suggested by our neural network ensemble (here we included training examples), and 5,538 genes representing the *S. cerevisiae* genome distribution.

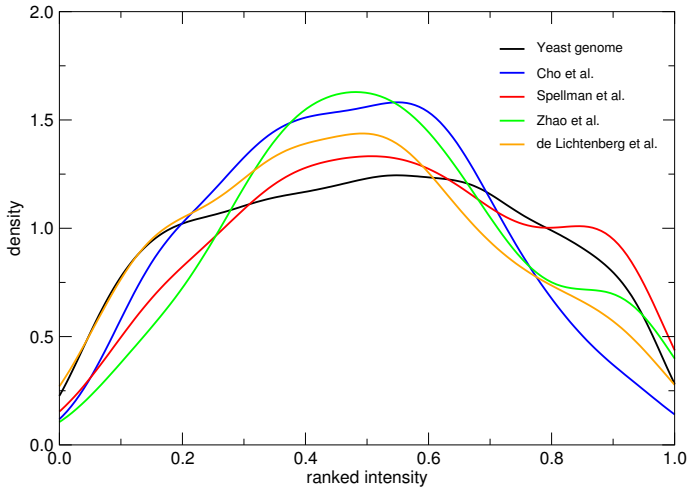


Figure 4: *The distribution of ranked intensities for different studies. Each distribution has been normalized to an area of one.*

Figure 4 demonstrates that all three sets of microarray identified genes contain fewer genes with low median intensity compared to both the genome distribution and our feature-based machine-learning method. These data even suggest that the fraction of weakly expressed genes identified in the studies drops with the stringency of the inclusion threshold, since the study that applies the most conservative inclusion criteria (Zhao *et al.*[3]) also identifies the lowest fraction of weakly expressed genes.

Temporal variation in protein features

To investigate temporal dynamics in *feature space* during the cell cycle we mapped the proteins identified with our method to time points in the cell cycle, based on the time of maximal expression of their encoding genes. The three publicly available cell cycle experiments (α-factor, $Cdc28$ and $Cdc15$) were used to determine the time of maximal expression of the identified cell cycle genes. The time series data was normalized within each experiment with the cycling times estimated by Zhao *et al.*[3] (58 min for the α-factor experiment, 115 min for the $Cdc15$ experiment and 85 min for the $Cdc28$ experiment) to bring the data on a comparable time scale. Within each experiment, the time of maximal expression was compared between two consecutive cycles, averaging the two time points if the time difference between them were less than 20% of the cell cycle period. In this way, a *peak time* was computed for the self-consistent genes in each experiment. If the gene did not meet these criteria, no value was computed for the gene in that experiment. The three experiments were then aligned by comparing the distribution of *peak times* for 46 genes known to peak in the G_1-phase[2], and furthermore shifted to set zero time to the suspected time of cell devision (G_1 entry). The three data transformations could be summarized as:

$$T_\alpha = \frac{t_\alpha}{58} - 0.146 \qquad T_{Cdc28} = \frac{t_{Cdc28}}{85} - 0.018 \qquad T_{Cdc15} = \frac{t_{Cdc15}}{115} - 0.099$$

where t_α is the number of minutes in the α-factor experiment, and T_α is time on the normalized and aligned timescale (between 0 and 1). The *peak time* thus indicates how many percent into the cell cycle a given cell cycle gene is maximally expressed. For every gene, the *peak times* were compared between those experiments where a value could be assigned (see above), and averaged only if the difference between them was less than 20% of the cell cycle period. Genes that did not meet this criteria were considered to show inconsistency in their expression and no final *average peak time* was reported for these genes. *Average peak time* assignment was attempted for the 500 highest scoring proteins identified by the neural network ensemble (including those of the training examples that score high). However, proteins encoded by genes displaying essentially no periodicity in any of the experiments (Fourier score below 1.5) were discarded to avoid applying the mapping procedure to noisy data. 309 of these 500 proteins met the criteria for consistency and periodicity and were assigned a unique, *averaged peak time* (based on one, two or three independent cell cycle experiments).

The cell cycle was divided into 100 time points (or percent) and the strength of a particular protein feature was calculated at each of the time points by averaging over the proteins expressed in a window of $\pm$ 5 time points. The strengths were visualized with respect to their deviation from the average value for all 309 periodically expressed proteins, using one color for values higher than the average and another color for lower values. The extremes of the color scale were set at $\pm$ two standard deviations. The temporal variation in the nine most relevant protein features is illustrated in Figure 5, where each circle corresponds to a feature. Zero time is at the top of the "clock", at the time of cell devision, i.e. entry into G_1 phase.

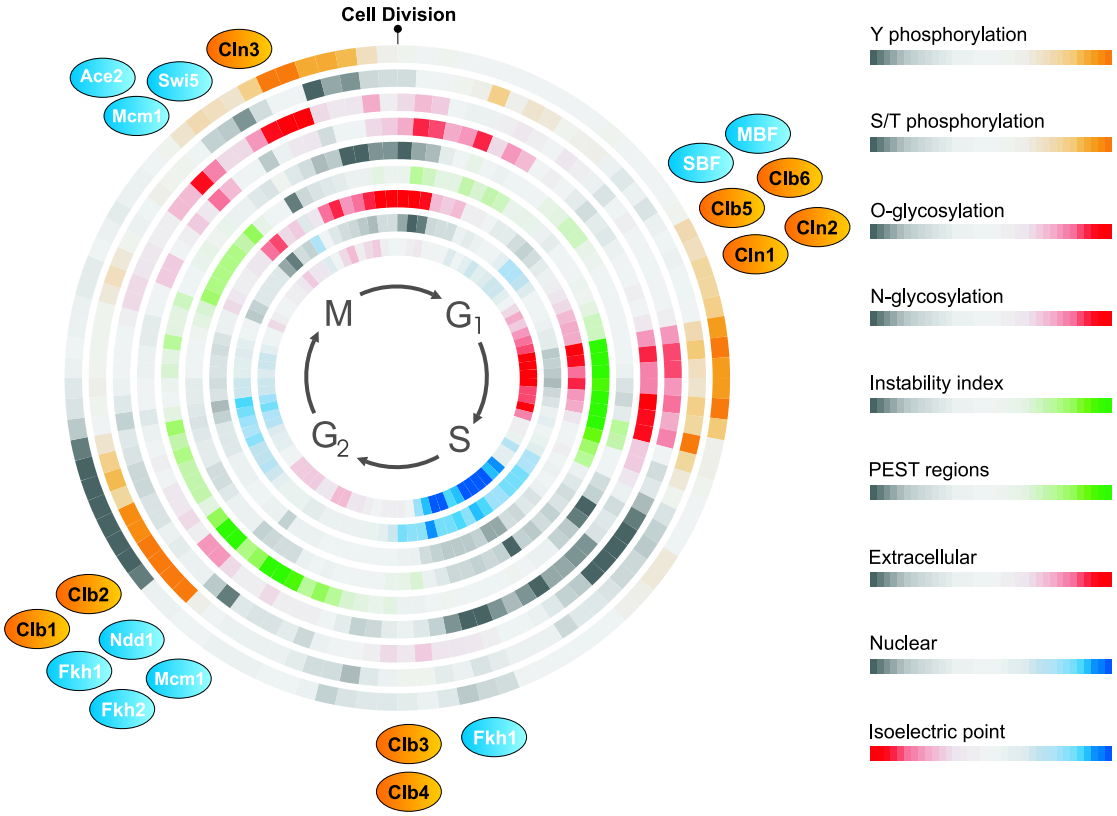


Figure 5: *Feature variation during the cell cycle. The temporal variation in nine selected protein features during the cell cycle, with zero time (at the top of the plot) corresponding to the presumed time of cell devision (M/G_1 transition). The color scales correspond to $\pm$ two standard deviations from the cell cycle average. The concentric feature circles correspond to: isoelectric point, nuclear and extracellular localization predictions[10], PEST regions[6], instability index[15], N-linked glycosylation potential, O-GalNAc glycosylation potential[8], serine/threonine- and tyrosine phosphorylation potential[5]. The presumed positions of the four cell cycle phases: G_1, S, G_2 and M are marked. Also depicted are known cell cycle transcriptional activators (marked in blue), positioned at the time where they are reported to function[16], along with nine cyclins (marked in orange), placed at the time where their genes are maximally expressed. Most of the cyclins are believed to activate Cdc28 kinase activity when expressed, but it should be noted that Clb5p and Clb6p are kept inactive in G_1 phase by the inhibitor protein Sic1p[17, 18].*

References

[1] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell,* **2**, 65–73.

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

[3] Zhao, L., Prentice, R. & 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.

[4] Skovgaard, M., Jensen, L.J., Brunal, S., Ussery, D. & Krogh, A. (2001). On the Total Number of Genes and Their Length Distribution in Complete Microbial Genomes. *TRENDS in Genetics*, **17 (8)**

[5] Blom, N., Gammeltoft, S., & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphsphorylation sites. *J. Mol. Biol.* **294**, 1351–1362.

[6] Rechsteiner, M. & Rogers, S. (1996). PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* **21**, 267–271.

[7] Nielsen, H., Brunak, S., Engelbrecht, J. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.

[8] Hansen, J., Lund, O., Tolstrup, N., Gooley, A., Williams, K. & Brunak, S. (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.* **15**, 115–130.

[9] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

[10] Nakai, K. & Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* **24**, 34–36.

[11] Mathews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta,* **405**, 442–451.

[12] Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A. & Brunak, S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265.

[13] Wood, V., Rutherford, K.M., Ivens, A., Rajandream, M.A. & Barrell, B. (2001). A Re-annotation of the *Saccharomyces cerevisiae* Genome. *Comp. Funct. Genom.* **2**, 143-154.

[14] Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielsen, H. B., Saxild, H., Nielsen, C., Brunak, S. & Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. *Gen. biol.,* **3(9)**, research0048.1–research0048.16.

[15] Guruprasad, K., Reddy, B. & Pandit, M. (1990). Correlation between stability of a protein and its di-peptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4**, 155–161.

[16] Simon, I., Barnett, J., Hannett, N., Harbison, C., Ranaldi, N., Volkert, T., Wyrick, J., Zeitlinger, J., Gifford, D., Jaakkola, T. & Young, R. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell,* **106**, 697–708.

[17] Mendenhall, M. & Hodge, A. (1998). Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **62**, 1191–1243.

[18] Breeden, L. (2000). Cyclin transcription: timing is everything. *Curr. Biol.* **10**, R586–R588.