doi:10.1016/S0022-2836(03)00490-X J. Mol. Biol. (2003) **329**, 663–674

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Protein Feature Based Identification of Cell Cycle Regulated Proteins in Yeast

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

Center for Biological Sequence Analysis, BioCentrum, The Technical University of Denmark, Building 208 DK-2800 Lyngby, Denmark

**Corresponding author*

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 have no known function.

Keywords: cell cycle; proteome dynamics; phosphorylation; subcellular location; instability/degradation

Introduction

The eukaryotic cell cycle is regulated at many levels, from transcription and translation to post-translational modification and targeted degradation. Its molecular machinery consists of highly diverse proteins, with little sequence similarity.[1–3] A major goal of cell cycle research is to uncover the size and complexity of the underlying molecular system, by identifying all cell cycle regulated genes and proteins.[4] Two DNA microarray studies have been performed in *Saccharomyces cerevisiae* in which expression levels for the yeast genome were measured during the cell cycle.[5,6] These data have been analyzed by visual inspection,[5] Fourier analysis and correlation to profiles of known cell cycle regulated genes,[6] as well as by a single-pulse statistical model.[7] Each study proposed a list of periodically expressed genes based on their analysis of the data. However, pronounced discrepancies exist between these lists of cell cycle regulated genes, as shown in Figure 1.

In these studies, a total of 940 genes were proposed to be periodic, yet less than half of them (397) were suggested by at least two groups and only 144 genes were identified by all three. Zhao *et al.*[7] analyzed the three cell cycle experiments (synchronized with α-factor, *Cdc*28 and *Cdc*15 temperature sensitive mutants) individually and concluded that 1088 genes showed significant periodicity in one of the experiments, 260 were periodic in at least two of three and only 71 genes were significant in all three experiments. This view is supported by a recent analysis by Shedden & Cooper which concludes that the periodicity of a given gene in one experiment (α-factor, *Cdc*28 or *Cdc*15) often cannot be reproduced in the other experiments.[8] Together, these observations demonstrate discrepancies both between the individual synchronization experiments analyzed with the same method, and between the conclusions of different research groups analyzing the same data set with different methods (Figure 1). A subset of the suggested periodic genes may thus be false positives. The variability does, presumably, not stem from the microarray technology as such, but is rather a result of different synchronization methods, different experimental conditions and different analysis methods. A genome-wide study was conducted recently to identify genes whose promoters are bound by one of nine known cell cycle transcription factors.[9] Interestingly, these authors could only detect binding to 27–50% of the 800 genes proposed by Spellman *et al.*[6]

The results presented here demonstrate that many cell cycle proteins display correlations between their features, which are different from

† U.L. and T.S.J. contributed equally to this work.

‡ http://www.cbs.dtu.dk

Abbreviations used: FHA, forkhead-association domain.

E-mail address of the corresponding author: brunak@cbs.dtu.dk

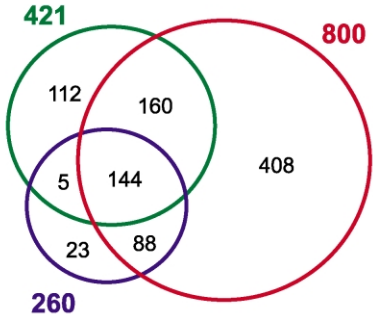

Figure 1. Extent of agreement between the published lists of periodically expressed transcripts. Data shown for Cho *et al*.[5] (green, 421 genes), Spellman *et al*.[6] (red, 800 genes) and Zhao *et al*.[7] (blue, 260 genes).

those of other yeast proteins. These features include phosphorylation, glycosylation, stability and/or disposition for targeted degradation, as well as localization in the cell. Further analysis reveals systematic temporal variations in the protein features during the cell cycle, demonstrating that many co-expressed cell cycle genes encode proteins that share the same features (even if they have no sequence similarity). Using such discriminative features as input, prediction algorithms were trained to identify cell cycle regulated proteins in the yeast proteome. Feature-based classification methods have been described for *Mycobacterium tuberculosis* and *Escherichia coli*,[33] and has been employed recently for function prediction of human proteins.[10] Our method identifies a set of new putative cell cycle regulated proteins.

Results

Following a new periodicity analysis of the publicly available microarray data, a training set was selected consisting of 97 proteins displaying very significant periodicity in expression during the cell cycle, along with 556 proteins encoded by non-periodic genes. Both sets display large diversity in biochemical function, sequence and three-dimensional structure. Neural networks were trained to distinguish cell cycle proteins from non-cell cycle proteins solely based on their features. Although *S. cerevisiae* was the first eukaryotic organism to be sequenced,[11] annotations of protein features are today only available for a subset of the proteins encoded in the genome. The classification method was therefore based on protein features calculated or predicted from the amino acid sequence by a set of well-documented bioinformatics tools and predictors (see Figure 2).

Characteristic features of cell cycle proteins

Clearly, not all features are of equal discriminative value for separating the two classes of yeast proteins. We therefore performed data-driven feature selection, where an iterative procedure

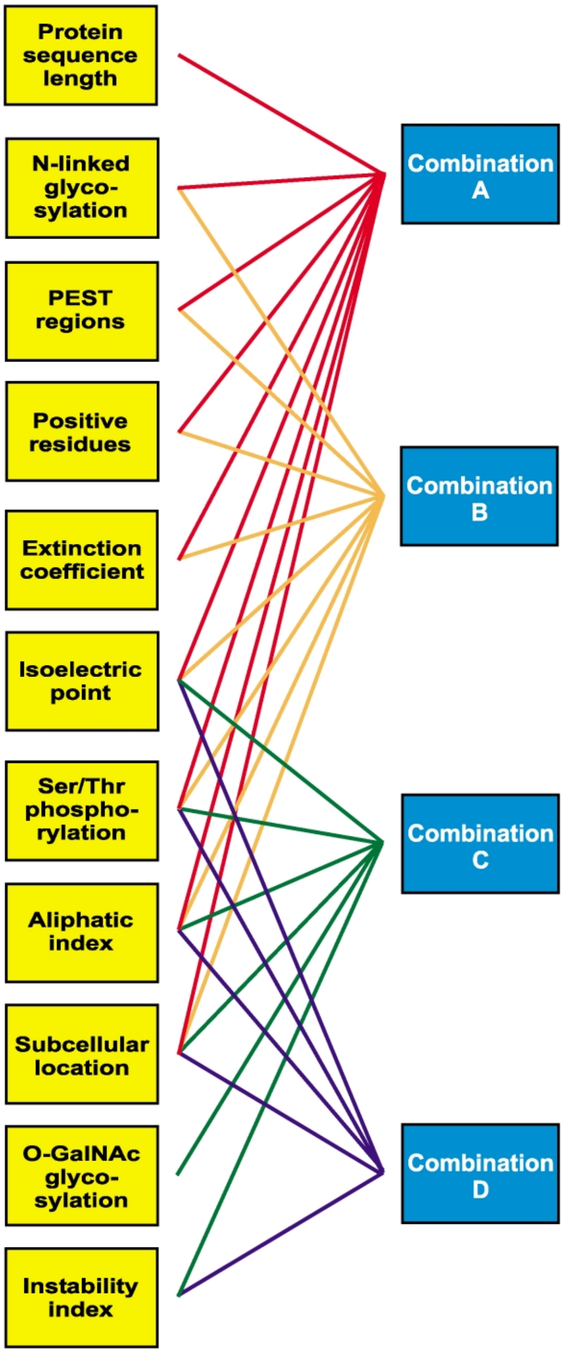

Figure 2. Characteristic protein features of cell cycle regulated proteins. The features selected by the neural networks for best discriminative value were: sequence length, N-linked glycosylation (Ramneek Gupta, unpublished results), PEST regions,[16] number of positively charged residues, extinction coefficient, isoelectric point, serine/threonine phosphorylation,[15] aliphatic index, subcellular localization,[25] O-GalNAc glycosylation[26] and instability index.[19] The different colored arcs show which features were used in each of the four input combinations. The following features were tested and discarded in the process due to their relatively poor discriminative value in input combinations: tyrosine phosphorylation,[15] signal peptides,[27] O-GlcNAc glycosylation (Ramneek Gupta, unpublished results), transmembrane helices,[28] hydrophaticity (GRAVY),[29] amino acid composition and number of negatively charged residues.

was used to select features that contribute the most to the predictive performance (see Materials and Methods). The iterative feature selection procedure started out with 18 features (Figure 2) and ended up with four powerful feature combinations, containing five to nine protein features. For increased predictive performance, the output from 15 neural networks using the four best input feature combinations were combined into one score. This ensemble integrated a total of 11 protein features (Figure 2) and had better predictive performance than any of the individual neural networks, when tested on independent test examples (see supplementary information†).

The discriminatory features provide an interesting characterization of cell cycle regulated proteins as a class, and many can be directly linked to existing knowledge of the cell cycle. Serine/threonine protein phosphorylation proved very useful for the classification. Our findings indicate that high potential for serine/threonine protein phosphorylation are over-represented in cell cycle regulated proteins, consistent with the known involvement of many serine/threonine kinases, e.g. the yeast Cdk, Cdc28p, in cell cycle regulation.[1] The predicted subcellular localization also proved very valuable for the discrimination. Cell cycle regulated proteins appear to be over-represented in the nuclear and cell wall categories, most likely explained by their involvement in processes such as transcription, DNA replication, repair, chromatin functions, budding and cell wall formation. Underrepresentation in other subcellular compartments is also very useful for the neural networks, since they are able to use negative information. Other correlations picked up by the neural networks indicate that many cell cycle proteins are unstable (have a high instability index) and/or contain so-called PEST regions in their amino acid sequence, regions known to be recognized by ubiquitin ligating complexes such as the anaphase promoting complex/cyclosome (APC/C) and the Skp1p-Cdc53p/Cullin-F-Box protein complexes (SCF) that target numerous cell cycle proteins for degradation by the proteasome.[3] Many cell cycle regulated proteins appear to have high potentials for N-linked glycosylation, a post-translational modification found almost exclusively in secreted or extracellular proteins, suggesting that many of these proteins could be related to budding and cell wall formation. All in all, the data-driven selection procedure identifies key features that are consistent with existing knowledge that phosphorylation, localization and degradation are major regulatory mechanisms of the cell cycle.[1–3]

Predictive performance of the neural network ensemble

The neural networks in the ensemble perform a complex integration of the feature information, and output a single numerical value, indicating to what degree a specific protein has combinations of features characteristic of cell cycle proteins. The predictive performance of the ensemble was estimated on the independent test sets (see Materials and Methods), and used to construct performance-curves showing sensitivity and false positive rate of the method as function of the threshold applied to the output from the neural network ensemble (Figure 3). Figure 3 shows how well the method works on test proteins (not used for training). At high threshold values the sensitivity is relatively low, but the method has also a very low false positive rate. This means that high scores are to be taken as strong supporting evidence for a cell cycle role, whereas low scores are less conclusive. In other words, the method will not identify all cell cycle proteins, rather it is suited for finding new putative candidates that may be missed by other techniques.

Proteome-wide prediction of cell cycle regulated proteins

The ensemble was applied to the *S. cerevisiae* proteome, to identify new putative cell cycle regulated proteins. No predictions were made on the proteins used for training. Two hundred and fifty proteins scored above a conservative threshold at 0.863, which in the performance-curve (Figure 3) corresponds to an estimated sensitivity of 37.8% and a false positive rate of 3.8%. From this list we removed proteins considered in a recent re-annotation by Wood *et al.*[12] to be “spurious” or “very hypothetical”, leaving a total of 211 proteins above the threshold (out of 5042 predictions). The method appeared to over-predict slightly on these ORFs, possibly because features such as PEST regions, glycosylation and phosphorylation were normalized with respect to the sequence length, meaning that very short sequences could appear relatively strong in these features. However, the average length of the highest scoring 211 proteins was essentially identical with the proteome average, meaning that the method does not prefer short sequences.

By assuming a given number of true cell cycle regulated proteins, N_{true}, among the total number of predictions, N_{total}, one may estimate the enrichment over a random predictor, E, by:

$$E = \frac{N_{\text{ensemble}}}{N_{\text{random}}} = \frac{\left[\dfrac{sN_{\text{true}}}{sN_{\text{true}} + f(N_{\text{total}} - N_{\text{true}})}M\right]}{\left[\dfrac{N_{\text{true}}}{N_{\text{total}}}M\right]}$$

$$= \frac{N_{\text{total}}}{N_{\text{true}} + \dfrac{f}{s}(N_{\text{total}} - N_{\text{true}})}$$

where s is the sensitivity, f is the false positive rate and M is the number of proteins included above

† http://www.cbs.dtu.dk/cellcycle

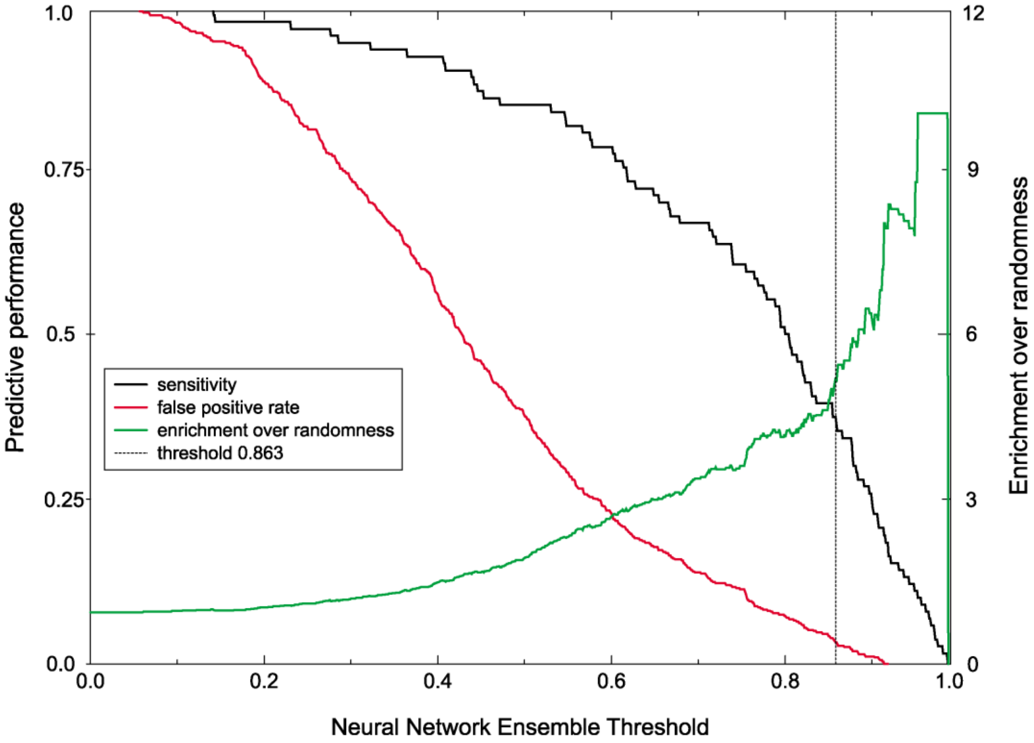

Figure 3. Predictive performance of the neural network ensemble. The false positive rate and sensitivity is plotted against the threshold applied to the ensemble output on the left axis, along with the estimated enrichment over a random predictor on the right. Here, we define the sensitivity as TP/(TP + FN) and the false positive rate as FP/(FP + TP), where TP are true positives (positive in reality, positive in prediction), FN are false negatives (positive in reality, predicted negative) and FP are false positives (negative in reality, predicted positive). The estimated enrichment was based on an assumption of 500 true cell cycle regulated proteins among the 5042 predicted, as described in the text.

the threshold. Assuming, $N_{\mathrm{true}} = 500$, the expected number of true hits should be $N_{\mathrm{ensemble}} = 110$ among the highest scoring 211 and $N_{\mathrm{random}} = 21$ by random, i.e. a fivefold enrichment. Figure 3 shows the estimated enrichment as a function of the applied threshold (assuming $N_{\mathrm{true}} = 500$). For comparison, $N_{\mathrm{true}} = 800$ and $N_{\mathrm{true}} = 300$ correspond to four- and 6.5 fold enrichment, respectively (at threshold 0.863). Another way to assess the predictive performance, is to consider the 104 known cell cycle regulated genes (compiled by Spellman *et al.*) and the 144 genes identified in all three microarray studies (see Figure 1).[5–7] Of these, 75 were used for training the neural network method, leaving 128 periodically expressed proteins, of which 19 were included among the 211 highest scoring. By random sampling $211(128/5042) = 5.3$ proteins should be expected, meaning a three- to fourfold enrichment. The estimated enrichment among the top-scoring 211 proteins thus depends somewhat on the analysis method, but is likely to be between three- and six-fold. The highest scoring 211 proteins and the proteome-wide predictions are available†.

Proteins predicted to be cell cycle regulated

Inspection of the predictions shows large diversity in function as well as subcellular location of the proteins suggested by the neural networks. Among the top-scoring proteins we found kinases, phosphatases, cyclins, transcription factors and proteins related to DNA replication/repair, cytokinesis, spindle pole body or cell wall biogenesis. The method, as expected from the statistics, also suggested proteins whose function appears unrelated to the cell cycle, such as mRNA splicing, intracellular transport or ribosomal proteins. However, many of the proteins identified have no known function, suggesting a high potential for new discoveries. Table 1 shows the 50 highest scoring proteins currently with unknown function in the *Saccharomyces* Genome Database (SGD).

The major strength of the feature-based approach is its independence of experimental errors and biases. It is itself not perfect, but it should not be expected to make the same types of errors as the methods it complements. Many of the highest scoring proteins have previously shown periodicity in one or more of the cell cycle gene expression experiments. In cases where the gene expression evidence is not consistent (periodic in some experiments and non-periodic in other), the neural network score provides an

† http://www.cbs.dtu.dk/cellcycle

Table 1. The 50 highest scoring protein with no annotated function in the *Saccharomyces* Genome Database

ORF	SGD name	ANN	Fourier	Zhao *et al.*[7]	Spellman *et al.*[6]	Cho *et al.*[5]
YIL169C		0.98	2.8			C
YDL038C		0.98	5.3			
YOR009W	TIR4	0.98	1.0			
YOL155C		0.97	3.0			C
YOL030W	GAS5	0.97	4.1	Z	S	
YOR220W		0.97	2.5			
YMR317W		0.97	2.1			
YLR194C		0.96	5.4		S	
YGR161C		0.96	2.4			
YFR054C		0.96	1.6			
YDL211C		0.96	3.3		S	
YHR212C		0.95	0.8			
YIR016W		0.95	0.8			
YDR451C	YHP1	0.95	5.3		S	C
YNL176C		0.94	3.7	Z	S	C
YMR233W	TRI1	0.94	0.8			
YNL074C	MLF3	0.94	0.8			
YLR040C		0.92	2.6		S	
YHR214W		0.92	2.3			
YAR066W		0.92	1.7			
YFR022W		0.92	2.2			
YPL014W		0.92	3.5		S	C
YPR171W		0.92	1.1			
YML041C	VPS71	0.92	0.9			
YGR035C		0.92	4.5		S	C
YNR014W		0.92	1.8			
YLR042C		0.91	1.6			
YNL078W	NIS1	0.91	5.3		S	C
YOR019W		0.91	0.8			
YBR162C	TOS1	0.91	2.6			
YDL129W		0.91	1.6			
YAL065C		0.90	2.0			
YBR203W		0.90	1.7			
YOL036W		0.90	1.0			
YDR223W		0.90	1.0			
YOR324C		0.90	3.3		S	
YJR115W		0.90	0.8			
YDR200C	VPS64	0.90	1.7			
YJL160C		0.89	2.0			
YGR079W		0.89	1.1			
YPL070W	MUK1	0.89	1.1			
YOL070C		0.88	3.4		S	C
YKR045C		0.88	1.0			
YDL037C		0.88	8.3		S	
YBR255W		0.88	1.3			
YLR003C		0.88	1.4			
YBL031W	SHE1	0.88	2.1			
YHR049C-A		0.88	1.1			
YLR031W		0.87	1.4			
YPL158C		0.87	6.1		S	C

ANN is the score from the artificial neural network ensemble, Fourier is the score from the Fourier periodicity analysis (see Materials and Methods) and Z, S, and C indicate whether the gene was included in any of the lists published by Zhao *et al.*,[7] Spellman *et al.*[6] and Cho *et al.*,[5] respectively.

independent source of supporting evidence. Interestingly, the ensemble also suggests a number of known cell cycle proteins (CDC24, SIR4, CDH1, MPS1, MLH2 and HPC2) that display no periodic expression. It is therefore possible, that the features are conserved among cell cycle proteins as such, and not just among those whose expression is periodic during the cycle.

The highest scoring of all proteins in the *S. cerevisiae* proteome is encoded by the gene *YIL169C*. It has no known function, but was proposed periodic by Cho *et al.*[5] This finding was, however, not confirmed by the lists published by Spellman *et al.*[6] and Zhao *et al.*[7] Figure 4 shows gene expression profiles for *YIL169C* and the two forkhead transcription factors *FKH1* and *FKH2*. These genes all display maximal expression approximately 50% into the cell cycle, in late *S* or early G_2 phase. *FKH1* and *FKH2* are known[2,9] to promote transcription of a large number of cell cycle genes in G_2/M, and recent genome-wide location data[9] suggest the promoter region of *YIL169C* to be associated with at least one of these known cell cycle transcriptional activator, Fkh2p, possibly also Fkh1p and Ndd1p. Furthermore, the protein product of YIL169C has reported protein–protein interactions with Mob1p and Fus3p.[13] Mob1p is required for cytokinesis and mitotic

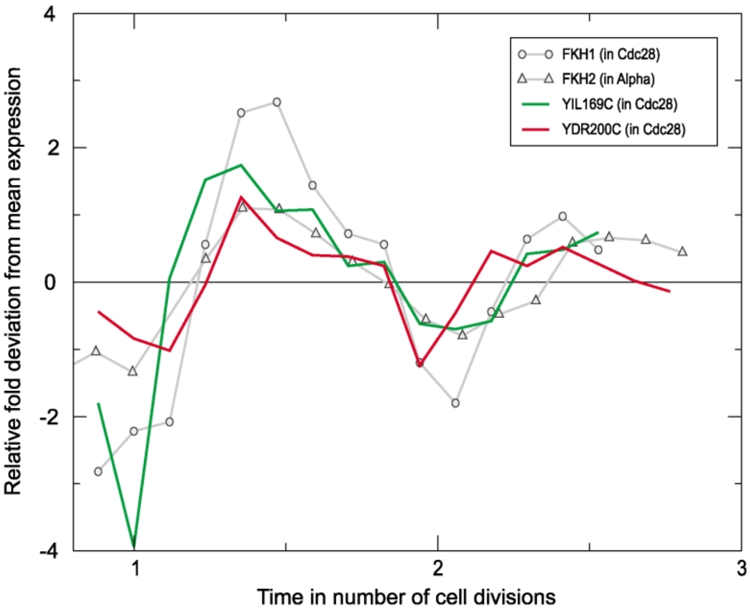

Figure 4. Gene expression profiles. Profiles of gene expression during the cell cycle for *FKH1*, *FKH2*, *YIL169C* and *YDR200C*. Data taken from the *Cdc*28-experiment of Cho *et al*.[5] and the α-factor experiment of Spellman *et al*.[6] The timescale was normalized and shifted to make the microarray data from different experiments comparable (see Materials and Methods). The first, second and third cell division are marked on the abscissa.

exit,[14] whereas the *FUS3* gene encodes a MAP serine/threonine kinase. These data indicate that YIL169C transcription could be activated in G_2/M phase, possibly by Fkh1p, and that the protein might play a role toward the end of the cell cycle. Prediction of phosphorylation sites indicates Yil169p to be heavily phosphorylated on serine and threonine residues.[15] Also, the sequence is predicted to contain several PEST regions,[16] indicative of a high potential for targeted degradation.

Both forkhead transcription factors contain a protein domain, the forkhead-association domain (FHA), demonstrated to specifically recognize and bind phosphothreonine epitopes on proteins.[17] Such domains are also present in the DNA damage checkpoint protein Rad53p and in the protein Tos4p, both of which are recognized by the neural network ensemble and found to peak at the presumed G_1/S transition. Interestingly, the neural network ensemble also identifies a protein, encoded by the gene *YDR200C*, which in one of the cell cycle experiments (the *Cdc*28 arrest) displays a cyclic pattern of expression (Figure 4) similar to that of *FKH1/2*, and which is also reported to contain an FHA domain.[17,18] It has no known function, but has reported interactions with another FHA containing protein of unknown function, Ylr238p, with Far3p, which plays a role in pheromone-mediated cell cycle arrest and with a protein of unknown function, Ynl127p, which shows weak similarity to Fus2p, a protein involved in cell fusion during mating. Our analysis finds *YDR200C* to be periodic, but only in the *Cdc*28-experiment performed by Cho *et al*.[5] It was, however, not included in any of the published lists of periodically expressed genes.[5–7] A possible explanation for this is the weak regulation of the gene (relative magnitude of up-regulation during the cell cycle was only around twofold).

Table 2 shows a selection of other proteins of unknown function that were suggested by the neural network ensemble, for which we have been able to find other sources of data that may support a cell cycle role. We show these to draw attention to what we believe to be some of the most promising and interesting candidates suggested by our method.

Weakly expressed genes

Figure 5 shows the distributions of signal intensity on microarrays (see Materials and Methods)

Table 2. Selected putative cell cycle proteins of unknown function, for which other sources of evidence exist that may support a role in the cell cycle

ORF	Other information
YJL160C	Mcm1/Swi5 binding, weakly expressed
YNL269W	Swi5 binding, weakly expressed
YBR162C/TOS1	Swi4 binding, SBF binding[31]
YAL028W	Fkh2 binding
YCL063W	Fkh1 binding, periodic in expression
YJL051W	Fkh2/Mcm1/Ndd1 binding, periodic in expression
YDR055W	Swi5 binding, periodic in expression
YJL078C/PRY3	Ace2/Swi5 binding, periodic in expression
YOL030W/GAS5	Fkh1 binding, periodic in expression
YLR194C	Swi5 binding, periodic in expression
YDR200C	FHA domain, possibly periodic in expression
YOR008C/SLG1	Cell cycle phenotype when overexpressed[32]

Data for binding of individual cell cycle transcription factors were taken from Simon *et al*.[9] and Lee *et al*.,[30] and only reported if the binding was significant ($p < 0.001$) in both studies. Data for SBF binding was taken from Iyer *et al*.[31] Periodicity in gene expression was based on the microarray data reported by Spellman *et al*.[6] "Cell cycle phenotype when overexpressed" refers to an over-expression screen performed by Stevenson *et al*.[32] that proposed a number of new cell cycle genes.

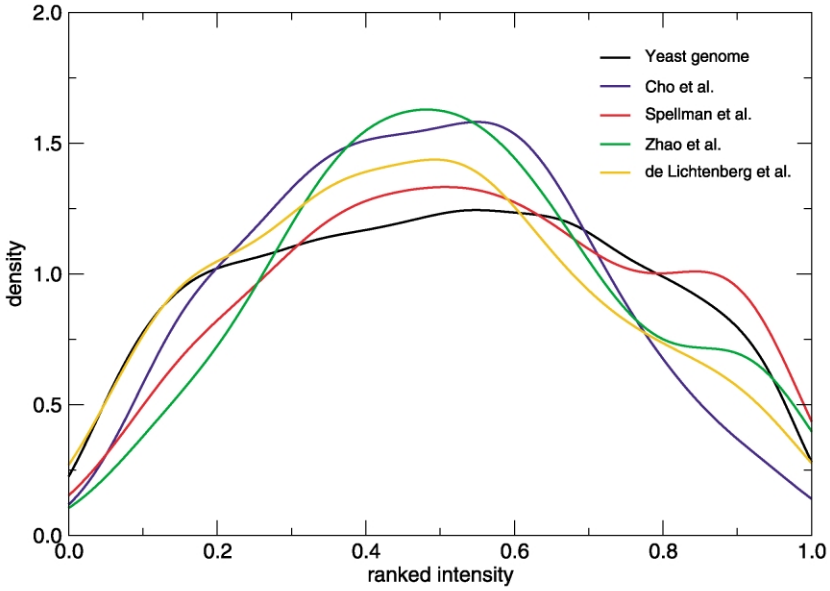

Figure 5. Distributions of ranked intensity for selected gene sets. The Figure shows smoothed distributions of median intensity (see Materials and Methods) for the entire *S. cerevisiae* genome, the cell cycle regulated genes proposed by Spellman *et al.*,[6] by Cho *et al.*,[5] by Zhao *et al.*[7] and the 500 highest scoring genes (no training examples) from the neural network ensemble. ORFs annotated as "spurious" or "very hypothetical" by Wood *et al.*[12] were removed from all sets.

for different sets of proposed cell cycle regulated genes ("spurious" and "very hypothetical" ORFs were removed from all sets). It demonstrates an under-representation of weakly expressed genes among the sets of genes identified as cell cycle regulated in microarray studies, compared to the entire genome distribution. The under-representation appears to be most significant for the genes identified by mathematical analysis of the data (Zhao *et al.*[7] and Spellman *et al.*[6]). The problem is most pronounced for the genes identified by Zhao *et al.*,[7] who used the most strict criteria and only included genes found significantly periodic in at least two of the three time series experiments (α-factor, *Cdc*15 and *Cdc*28). The under-representation thus appears to be largest among the genes for which the data speaks most convincingly for a periodic expression. A likely explanation for this is that weakly expressed genes have a poorer signal-to-noise ratio on microarrays. In comparison, the neural network method does not display any significant under-representation (Figure 5), and may therefore identify cell cycle regulated genes previously undetected on microarrays.

Temporal feature variation of the cell cycle proteome subset

To assess temporal variations in protein feature space, the 500 highest scoring proteins from the predictions were mapped to time-points in the cell cycle, based on the time of maximal expression of their encoding genes in the three gene expression experiments.[6] Here, we included the periodically expressed genes used for training the method. Of these 500 genes, 309 could be confidently assigned a time of maximal expression (see Materials and Methods). The strength of a particular feature (e.g. isoelectric point) was computed at all time-points during the cell cycle by averaging over proteins whose encoding genes peak in the neighborhood of the particular time-point. This yielded a division of the cell cycle into a "clock", with each time-point corresponding to 1% of the cell cycle. Zero time was set to be the presumed position of G_1 phase entry, right after cell division. A circular plot was constructed (Figure 6) where each of the concentric circles corresponds to a particular feature. The color indicates whether proteins expressed around a given time-point have higher or lower value of a given feature than the average of all the 309 mapped proteins.

Figure 6 thus offers a novel *in silico* view of the cell cycle proteome dynamics and reveals intriguing temporal variations in characteristic features of these proteins during the cell cycle. The time resolution of this "clock" is much higher than the conventional division of the cycle into four phases (G_1, S, G_2 and M) depicted inside the feature circles of Figure 6. Our results demonstrate that genes maximally expressed at the same stage in the cell cycle appear to share features at the protein level. The patterns observed in Figure 6 were largely conserved in similar plots representing the sets of periodically expressed genes previously suggested in microarray studies[5–7] (data not shown), suggesting that the feature patterns of Figure 6 may be representative of the entire yeast cell cycle proteome subset. Also depicted in Figure 6 are known cell cycle transcriptional activators (marked in blue) positioned at the time where they are reported to function,[9] along with nine cyclins (marked in orange) placed at the time where their genes are maximally expressed.

Feature patterns during the cell cycle

In the beginning of the cell cycle (going clockwise from the top of Figure 6), in early G_1, a large uniformly colored area was observed, which indicates that no features are over or underrepresented among the proteins expressed here. This changes

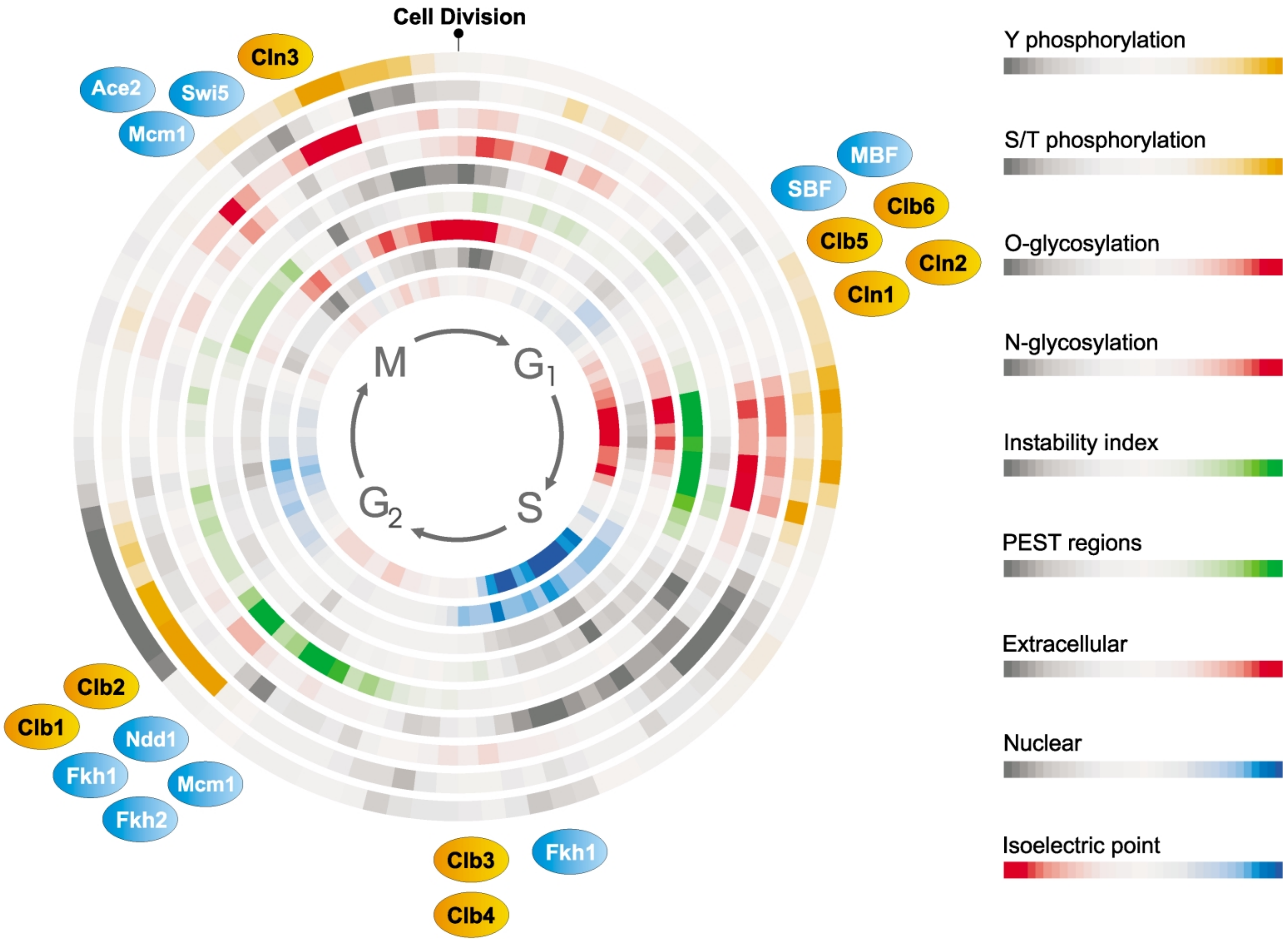

Figure 6. Feature variation during the cell cycle. The temporal variation in nine selected protein features during the cell cycle, with zero time (at the top of the plot) corresponding to the presumed time of cell division (M/G_1 transition). The color scales correspond to ± two standard deviations from the cell cycle average. The concentric feature circles correspond to: isoelectric point, nuclear and extracellular localization predictions,[25] PEST regions,[16] instability index,[19] N-linked glycosylation potential, O-GalNAc glycosylation potential,[26] serine/threonine phosphorylation potential[15] and tyrosine phosphorylation potential.[15] 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,[9]

into a very distinct feature pattern 20–30% into the cycle (late G_1), where proteins have a large number of features in common, particularly post-translational modifications and PEST regions. This pattern correlates with the maximal expression of several cyclins (*CLN1*, *CLN2*, *CLB5*, *CLB6*), with the expression of genes involved in budding (*BUD9*, *BNI4*, *GIN4*, *CSI2*, *CRH1*, *AXL2*, *SVS1*, *QRI1*, *MCD4*, *RSR1*, *MSB2*, *MNN1*) and DNA replication/repair (*PRI2*, *DPB2*, *POL12*, *POL30*, *DBP2*, *CDC9*, *CDC21*, *RAD53*, *MSH6*, *RFA1*, *RAD27*, *RNR1*, *SLD2*, *CTF18*, *TOF1*, *RFA2*, *OGG1*, *CDC45*, *HYS2*, *MSH2*, *POL2*), as well as with the timing of MBF and SBF activity. Together, these data suggest the feature space pattern 20–30% into the cell cycle to be a fingerprint of the G_1/S transition (termed START in yeast). Figure 6 indicates that this group of G_1/S proteins contains many members with PEST regions, relatively low isoelectric point, high potential for phosphorylation and glycosylation. This group of proteins is also predicted to be rich in extracellular proteins or proteins related to the cell wall. The latter correlates well with proteins involved in budding and cell wall formation.

The late G_1 pattern changes completely 35–45% into the cycle (Figure 6) caused by the expression of a new group of proteins characteristic of being mostly nuclear, having very high isoelectric points, being very stable (low instability index and few PEST regions) and displaying lower potential for glycosylation and phosphorylation than the average yeast cell cycle regulated protein. Among the proteins expressed here are eight histones (Hht1p, Htb1p, Htb2p, Hhf1p, Hhf2p, Hht2p, Hta1p and Hho1p) supporting the notion that this pattern corresponds to the S-phase of the cell cycle. Histones are stable, nuclear proteins with high isoelectric point and no potential for glycosylation. But the histones are not the only proteins in this group that possess these characteristic features. *IRS4*, *SHE1*, *TOF2*, *ENT4* and *YNR014W* all encode proteins with high isoelectric point (all above 8.0), and are all predicted to be nuclear. Most of these have no reported relation to the cell cycle.

Only few features stand out in the presumed G_2 phase. The proteins expressed appear to have a higher instability index, indicating a short lifetime. Unlike PEST regions, this feature does not directly imply targeted degradation of the proteins. Rather it has been found that certain protein compositions render the proteins unstable.[19] Almost as a burst, the serine and threonine phosphorylation increases in a small window in G_2, where the tyrosine phosphorylation potential at the same time reaches its lowest level throughout the cycle. It coincides with the maximal expression of the two cyclin genes *CLB1* and *CLB2*, and with the reported function of the transcriptional activators Mcm1p, Fkh1p, Fkh2p and Ndd1p that activate the transcription of G_2/M genes.[2,9]

Towards the end of the cell cycle, in mitosis, Figure 6 displays a very complex pattern of feature strengths, which indicates that many subgroups of proteins with distinct features are expressed here. In general, there are relatively few nuclear proteins and relatively many extracellular proteins. Glycosylation potentials are high, with the O-glycosylation most abundant in the early stage and the N-glycosylation dominating in the later proteins. Tyrosine phosphorylation is strong, whereas the serine/threonine phosphorylation potential is relatively low, in good agreement with existing knowledge that the kinase activity decreases at this stage.[1,20] PEST regions are abundant in some subgroups, whereas late M proteins appear to be relatively stable (low instability index). The complex patterns may result from the fact that the M phase is composed of sub-phases (prophase, metaphase, anaphase and telophase), where the subphase proteins represent different combinations of features.

Protein phosphorylation during the cell cycle

We find the changing patterns of phosphorylation particularly interesting. In general, our results indicate that serine/threonine phosphorylation is highly over-represented in cell cycle proteins. However, Figure 6 demonstrates significant temporal variations in both kinds of phosphorylation at three stages in the cycle, all with a different correlation between the two types. Proteins mapped to time-points 20–30% into the cell cycle have high potentials for both kinds of phosphorylation with the tyrosine potential rising first. The next differential phosphorylation pattern is seen 60–70% into the cell cycle, where proteins have high potentials for serine/threonine phosphorylation, but very low potentials for tyrosine phosphorylation. Towards the end of the cell cycle, before cell division, tyrosine phosphorylation peaks again, whereas the serine/threonine phosphorylation reaches its lowest level. These observations are, at least in part, consistent with previous experimental observations that the activity of the cyclin dependent kinases rise during the cell cycle from the G_1/S transition (START) until the end of mitosis, where it drops due to activity of inhibitors and targeted degradation of the cyclins.[1,20] Our results also suggest tyrosine phosphorylation to be more abundant among proteins expressed both at the suspected G_1/S

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.[1,2]

transition and towards the end of the cell cycle in mitosis. The literature describes several examples of tyrosine phosphorylation related to the cell cycle. Swe1p and Mih1p regulate phosphorylation of Tyr19 in Cdc28p, thereby controlling the timing of nuclear division[21] and entry into mitosis.[1] Swe1p is a tyrosine kinase whose transcription peaks in late G_1. Other cell cycle related tyrosine kinases are Rad53p (involved in DNA replication and DNA damage checkpoints) and Mps1p (involved in spindle pole body duplication and the spindle checkpoint in M phase). Furthermore, it has been discovered that nuclear localization of Cdc48p in late G_1 is controlled by tyrosine phosphorylation of the protein.[22] These examples confirm that tyrosine phosphorylation plays a role in late G_1/early S phase and in mitosis. The feature pattern (Figure 6) indicates that there may be many other instances of cell cycle related tyrosine phosphorylation not yet discovered.

Link between protein features and function

The classification approach used in this study relies on the conservation of protein features within a class of proteins, namely those involved in the cell cycle. A similar approach, the ProtFun method,[10] has recently been developed for predicting other functional categories of human proteins. The most recent analysis of this work reveals that features are more conserved among orthologs than paralogs, indicating that protein features are selectively conserved among proteins with similar function (L.J.J. *et al.*, unpublished results). This also seems reasonable, since many functionalities require the presence and recognition of short sequence motifs for post-translational modification or binding of other factors to the protein, but do not require the sequence or structure to be conserved. Proteins with the same cellular role will thus often be similar in protein feature space, but not necessarily similar at the level of protein sequence or three-dimensional structure. Cell cycle regulated proteins constitute a very broad class, but it would still be expected that certain functionalities or features are characteristic, if not unique to this class, distinguishing them from other proteins in feature space. Interestingly, we have found the temporal variations in protein features (Figure 6) to be largely conserved between the subsets of proteins identified by the neural network and those identified in microarray studies. Since the two identification approaches are independent, this suggests the characteristic features identified in this study and their dynamics to be generalizable to the entire yeast cell cycle proteome subset.

Cell cycle proteins in the human proteome

The ProtFun method[10] was trained on human data alone, but recent, unpublished data shows that it works surprisingly well on other eukaryotic organisms, including yeast. Inspired by this, we have applied the cell cycle prediction method trained on yeast to a set of protein sequences corresponding to all predicted genes in the human genome, as identified in the Ensembl database.[23] The data are not fully analyzed yet, but among the top scoring candidates with known function are several histones, the cell division kinase Cdk5p, the Cdk inhibitor Cdkn3p, the T1 cyclin, the DNA replication proteins Mssp1p and Mcm3p, the p53-binding protein Mdm4p and Rbbp8p which interacts with both Brca1p and the retinoblastoma protein. Another alternative in identifying human cell cycle proteins is to use the recent human cell cycle gene expression data and apply a method similar to that described here for yeast. Work on these data is in progress.

Materials and Methods

Training set

A periodicity analysis was performed on the three publicly available synchronization experiments (α-factor, *Cdc*28 and *Cdc*15) compiled by Spellman *et al.*[5,6] to identify periodically as well as non-periodically expressed genes in *S. cerevisiae*. A Fourier-like analysis was applied to the data, such that each gene i was assigned a score D_i based on its temporal expression profile during the experiment, with cell cycle frequency $\omega = 2\pi/T$:

$$D_i = \sqrt{\left(\sum_t \sin(\omega t)x_i(t)\right)^2 + \left(\sum_t \cos(\omega t)x_i(t)\right)^2}$$

The cell cycle periods, T, estimated by Zhao *et al.*[7] were used (58 minutes for the α-factor experiment, 115 minutes for the *Cdc*15 experiment and 85 minutes for the *Cdc*28 experiment), and a combined Fourier score, F_i, was computed as:

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

The contribution from the *Cdc*15 experiment was scaled in the combined score, because this experiment covers 2.5 cell cycles, whereas the α-factor and *Cdc*28 experiments cover only two (using the Zhao *et al.*[7] estimates). The lowest scoring 556 genes (threshold at 0.75) were used as examples of "non-cell cycle regulated proteins", which display no periodic regulation during the cell cycle. By an estimation method described in detail elsewhere†, we selected 115 genes in a set of very significantly periodic genes (a conservative threshold at 6.0). To ensure consistent behavior over multiple cycles, we required the Pearson correlation between the expression profiles of the first and the second cycle to be above 0.4, thereby excluding 18 genes. The procedures outlined above resulted in a training set consisting of 97 "cell cycle regulated proteins" and 556 "non-cell cycle regulated proteins".

† http://www.cbs.dtu.dk/cellcycle

Neural network training

Threefold cross-validation was used (division of the data set in three different ways), each with 430 protein sequences for training and 215 for independent evaluation of the classification performance, which was measured as the Matthews test correlation coefficient over all three test sets.[24] As input to neural networks we used protein features derived directly from the amino acid sequence of the proteins to give a feature space representation of each protein.

An iterative heuristic (similar to that used by Jensen *et al.*[10]) was applied to select the most discriminative features and the best performing combinations of these: combinations of two features were tested to select the best performing pairs, from which combinations of three features were generated. The best performing of these were used as starting points for generating combinations of four features, etc. The method thus only retains features that perform well alone or in combination with one other feature. In that way, irrelevant features are filtered out in the very beginning of the selection procedure.

For each input combination the performance was measured as the combined Matthews correlation coefficient[24] over all three independent test sets (for further details, see supplementary information†). Eighteen features were investigated and Figure 2 shows the four best input combinations, which were used in an ensemble of neural networks. For rescaling the individual network output, the test set scores were ranked and their distribution used as conversion table for output from that network, making it possible to average all 15 neural network output scores into one final score (between 0 and 1). As expected, the ensemble outperformed all of the individual neural networks. The performance of the ensemble was only tested on sequences not used for training. A prediction was obtained from the trained ensemble for the entire *S. cerevisiae* proteome (set of all translated ORFs from SGD‡). No predictions were, however, made for the training examples and the "spurious" or "very hypothetical" ORFs described by Wood *et al.*[12]

Microarray intensity distributions

The median fluorescence intensity (microarray spot intensity) was computed for each gene in each of the three cell cycle time series experiments[5,6] (α-factor, *Cdc*28 and *Cdc*15). Within each experiment, the median intensities were ranked. The median rank over all three experiments was then used as measure of the median intensity of each gene. For additional details, see the website.

Temporal variation in protein features

The three cell cycle experiments[5,6] (α-factor, *Cdc*28 and *Cdc*15) were used to determine the time of maximal expression of periodic cell cycle genes. The time series data were normalized within each experiment with the cycling times estimated by Zhao *et al.*[7] 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 was less than 20% of the cell cycle period. In this way, a peak time was computed only for the self-consistent genes in each experiment. The three experiments were aligned by comparing the distribution of peak times for genes known to peak in the G_1-phase,[6] and furthermore shifted to set zero time to the suspected time of cell division (beginning of G_1). The peak time thus indicates how many percent into the cell cycle a given cell cycle protein is maximally expressed. The peak times were compared between the three experiments and averaged only if the difference between them was less than 20% of the cell cycle period. Out of the 500 top-scoring proteins, 309 met this double self-consistency criterion and were assigned a unique average peak time (based on one, two or three experiments).

The cell cycle was divided into 100 time-points and the strength of a particular protein feature was calculated at each of the time-points by averaging over the proteins with average peak time in a window of ±5 time-points. The strengths were visualized with respect to their deviation from the average value of all 309 cell cycle regulated 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 ±2 standard deviations. The temporal variation in the nine most interesting protein features is illustrated in Figure 6.

† http://www.cbs.dtu.dk/cellcycle

‡ http://genome-www.stanford.edu/Saccharomyces/

Acknowledgements

The authors thank Jiri Bartek for valuable comments on this work, as well as Hans Henrik Stærfeldt for developing the Genewiz software used to construct Figure 6. This work was supported by grants from Novo Nordisk and the Danish National Research Foundation.

References

1. 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.
2. Breeden, L. (2000). Cyclin transcription: timing is everything. *Curr. Biol.* **10**, R586–R588.
3. Tyers, M. & Jorgensen, P. (2000). Proteolysis and the cell cycle: with this RING I do thee destroy. *Curr. Opin. Genet. Dev.* **10**, 54–64.
4. Nurse, P. (2000). A long twentieth century of the cell cycle and beyond. *Cell*, **100**, 71–78.
5. Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L. *et al.* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
6. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M. *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
7. Zhao, L., Prentice, R. & Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.

8. Shedden, K. & Cooper, S. (2002). Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucl. Acids Res.* **30**, 2920–2929.
9. Simon, I., Barnett, J., Hannett, N., Harbison, C., Ranaldi, N., Volkert, T. *et al.* (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.
10. Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C. *et al.* (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265.
11. Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L. *et al.* (1997). The yeast genome directory. *Nature*, **387**, 5–105.
12. Wood, V., Rutherford, K., Ivens, A., Rajandream, M. & Barrell, B. (2001). A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2**, 143–154.
13. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
14. Luca, F., Mody, M., Kurischko, C., Roof, D., Giddings, T. & Winey, M. (2001). *Saccharomyces cerevisiae* mob1p is required for cytokinesis and mitotic exit. *Mol. Cell. Biol.* **21**, 6972–6983.
15. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362.
16. Rechsteiner, M. & Rogers, S. (1996). PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* **21**, 267–271.
17. Durocher, D. & Jackson, S. (2002). The FHA domain. *FEBS Letters*, **513**, 58–66.
18. Letunic, I., Goodstadt, L., Dickens, N., Doerks, T., Schultz, J., Mott, R. *et al.* (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**, 242–244.
19. 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.
20. Lew, D., Weinert, T. & Pringle, J. (1997). Cell cycle control in *Saccharomyces cerevisiae.* In *The Molecular Biology of the Yeast* Saccharomyces*-3: Cell Cycle and Cell Biology* (Pringle, J., Broach, J. & Jones, E., eds), pp. 607–695, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
21. Murakami, H. & Nurse, P. (2000). DNA replication and damage checkpoints and meiotic cell cycle controls in the fission and budding yeasts. *Biochem. J.* **349**, 1–12.
22. Madeo, F., Schlauer, J., Zischka, H., Mecke, D. & Frhlich, K. (1998). Tyrosine phosphorylation regulates cell cycle dependent nuclear localization of cdc48p. *Mol. Biol. Cell*, **9**, 131–141.
23. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L. *et al.* (2002). The Ensembl genome database. *Nucl. Acids Res.* **30**, 38–41.
24. Mathews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
25. 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.
26. 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.
27. 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.
28. 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.
29. Kyte, J. & Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
30. Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G. *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science*, **298**, 799–804.
31. Iyer, V., Horak, C., Scafe, C., Botstein, D., Snyder, M. & Brown, P. (2001). Genomics binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
32. Stevenson, L., Kennedy, B. & Harlow, E. (2001). A large-scale overexpression screen in *Saccharomyces cerevisiae* identifies previously uncharacterized cell cycle genes. *Proc. Natl Acad. Sci. USA*, **98**, 3946–3951.
33. King, R. D., Karwath, A., Clare, A. & Dehapse, L. (2000). Accurate prediction of protein functional class in the *M. tuberculosis* and *E. coli* genomes using data mining. *Yeast* **17** (*Comp. Funct. Genom.* **1**), 283–293.

Edited by S. Reed

(Received 13 September 2002; received in revised form 3 February 2003; accepted 8 April 2003)