Comparative Analysis of Cell Cycle Regulated Genes in Eukaryotes

Jerzy Dyczkowski dyczkows@molgen.mpg.de
Martin Vingron vingron@molgen.mpg.de

Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

Abstract

We compared microarray experiments on cell cycle of three model eukaryotes: budding and fission yeast and human cells. Only 112 orthologous groups were cyclic in the three model organisms. The common set of cyclic orthologs includes many taking part in the cell cycle progression, like cyclin B homologs, CDC5, SCH9, DSK2, ZPR1. Proteins involved in DNA replication included histones, some checkpoint kinases and some proteins regulating DNA damage and repair. Conserved cyclic proteins involved in cytokinesis included myosins and kinesins. Many groups of genes related to translation and other metabolic processes were also cyclic in all three organisms. This reflects rebuilding of cellular components after the replication and changes of metabolism during the cell cycle. Many genes important in cell cycle control are not cyclic or not conserved. This includes transcription factors implicated in the regulation of budding yeast cell cycle. The partially overlapping roles of regulatory proteins might allow the evolutionary substitution of components of cell cycle.

Keywords: cell cycle, cell cycle control, microarray analysis, evolutionary conservation

1 Introduction

The cell cycle is a fundamental biological process and its phases and many subcellular processes are present in all eukaryotes. DNA microarray studies provided a new tool to study this process. Several model organisms were studied in this way [8, 9, 19], and for some of them more than one study is available [2, 11, 13, 16]. So far only one study attempted to compare cell cycle microarray experiments across several eukaryotes [10].

There was a discussion on the validity of cell cycle microarray experiments because of the effect of synchronization methods [3, 17]. As one remedy, bioinformatics methods have been developed to determine periodically expressed genes even under the assumption of different phasing among the studied cells [7, 20] this study we utilize one of these methods [20].

Genetic mechanisms of cell cycle control might be expected to be conserved in eukaryotes. Therefore the discovery that only about 40-80 genes are cyclic both in budding and fission yeast [13] was met with surprise and generated interest in identifying the most important elements of cell cycle in all eukaryotes [15]. In this paper we study cell cycle microarray experiments on three model eukaryotes: *Saccharomyces cereviscae* [2, 16], *Schizosaccharomyces pombe* [13] and *Homo sapiens* HeLa cells [19]. We establish genes which cyclic expression was statistically significant and a subset of cyclic genes with orthologs in all three model organisms. We use them to establish which events during the cell cycle are conserved in all eukaryotes.

2 Material and Methods

We compared datasets of microarray experiments on cell cycle of *Saccharomyces cereviscae* [2, 16], *Schizosaccharomyces pombe* [13] and *Homo sapiens* HeLa cells [19]. Published experiments on *Arabidopsis thaliana* cells contained data on only ca. 14,200 genes [8] and 87 genes [9], of ca. 26,000 protein coding genes in plant genome. The second study researched only homologs of several cell cycle genes. For this reason we did not include Arabidopsis in this study.

The significance of cyclic changes in gene expression was verified by the method based on so-called average periodograms and g-statistic test [20] implemented in the package GeneTS for R. This method first computes the average periodogram, which is a graphical representation of presence of periodic transcripts in the data. This is a simple extension of standard periodogram, a tool widely used in analysis of time series. Next, it computes an exact statistic test to identify the subset of genes, which are actually periodic and quantify the results obtained by the average periodogram. This exact test is based on Fisher's g-statistic. This allows distinguishing periodic from randomly variable genes. The test of significance is carried on all genes simultaneously using the false discovery rate for multiple testing. The authors ran their methods on several periodic data sets, reporting the number of identified cyclic genes. We used a false discovery rate of $q<0.05$ as statistically significant.

Advantages of this method include that it checks for periodicity but makes only minimal other assumption about the shape of expression changes. Other methods, e.g. [7], typically assume a certain shape of expression curve (e.g. sinusoid) and check genes for similarity to that. In the case of *S. pombe*, where eight separate experiments were concluded, only genes found significantly periodic in at least three of experiments were counted. In this way we established a set of *cyclic genes* in all organisms.

Orthologs were defined according to the COG database [18]. We identified *cyclic conserved genes* as genes which were cyclic and belonged to an orthologous group containing at least one cyclic gene from *S. cereviscae, S. pombe* and *H. sapiens* each. Several genes in one organism can belong to the same orthologous group. Several large groups were searched for subgroups using the in-paranoid database [12].

To get an overview of expression changes of genes we used correspondence analysis of cyclic genes, as explained in [4, 5]. This is a graphical method, which starts with a matrix of genes versus experiments and computes the principal components of variability within the matrix. Experiments and genes are then plotted in a low dimensional space, typically a plane, according to the principal components identified in analysis. Genes that are upregulated in particular conditions tend to appear in the same direction from the center as the experiments in which they are upregulated. In our application, this allows the user to spot which cyclic gene peaks in which phase(s) of the cell cycle. Within such a plot one can further highlight additional features, like in our case evolutionary conservation.

3 Results and Discussion

We found 1296 genes in *S. cereviscae*, 1783 in *S. pombe* and probes corresponding to ca. 7567 genes in *H. sapiens* to be significantly cyclically expressed. These numbers of genes are much higher than those considered by the respective authors of the original microarray experiments. However, these authors used very conservative cutoffs in order to focus on the genes driving the cell cycle. Sherlock [15] states that the figures in the original papers are likely underestimates with respect to all genes involved in cell division. As stated in Methods, our numbers were determined by the method of [20] who also arrive at similar counts to the ones given here. While including only genes that cycle significantly according to these authors, our set contains also lowly expressed genes.

One previous study [10] compared cell cycle experiments on budding yeast, human cells and Arabidopsis cells. Because, as previously mentioned, the only available Arabidopsis data contains only about half of genes present in its genome, many conserved genes were likely missed and the true

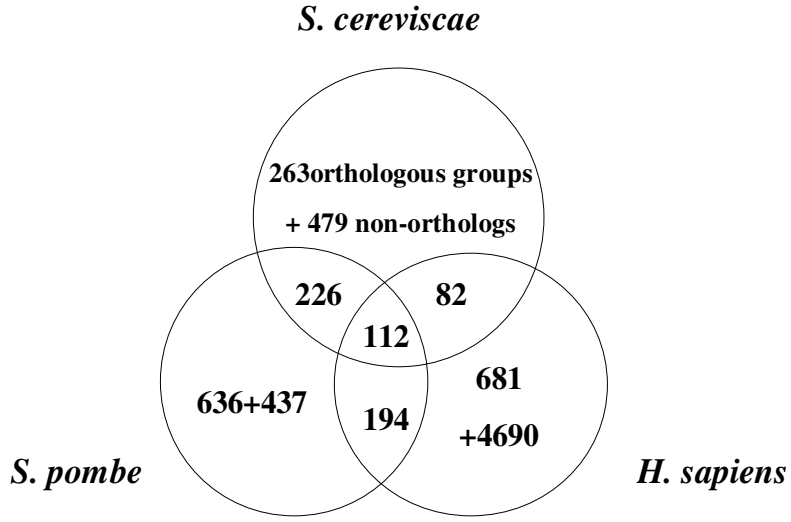

Figure 1: Overview of genes which oscillate during the cell cycle and have or lack orthologs on three model eukaryotes.

number of conserved genes was impossible to establish. The previous study also contains a brief discussion. Comparing to this study, we had the possibility of including a third complete genome and establishing the number of conserved and unconserved genes between the respective genomes. We also tested objectively how conserved the cell cycle is and looked detaily at the particular groups of genes.

Only a minority of cyclic genes belonged to orthologous groups shared between all organisms - only 112 orthologous groups were identified (Figure 1). We will refer to these as conserved cyclic genes. Those include 14 of 36 groups of orthologs found cyclic in two yeast species by Rustici *et al.* [13]. Among the remaining groups cited by [13], 8 have no ortholog in *H. sapiens* according to COG, 12 have not cyclic orthologs in *H. sapiens* and two were refound as significantly cyclic in this study.

The number of 112 conserved cyclic orthologs may appear low. We conducted a simulation to determine whether this number is lower or higher than expected in such group of genes. We emulated a situation, when one randomly draws some genes from the three genomes and checks how many orthologous groups were picked between them. For each organism under consideration there is a particular number of cyclic genes. The concrete figures were 1296 in budding yeast genes, 1783 in fission yeast genes, and 7567 human genes. The question is how many of those fall into orthologous groups according to COG. Thus, we repeatedly selected 1296, 1783, and 7567 genes, respectively, and counted how many orthologous groups these would belong to. Based on 100,000 randomizations, we found that 76% of these runs resulted in a count below 112 (mean=106,88, SD=8,86). This demonstrates that we are looking at a fairly typical number.

We further computed a correspondence analysis plot of the expression of cyclic genes in the human cell cycle (Figure 2). Blue dots correspond to genes, hybridisations are indicated by the cell cycle phase from which they were taken, and black crosses refer to conserved cyclic genes. One observes a prominent cluster of conserved cyclic genes at the bottom of the cloud of points and another, less pronounced cluster above the cloud. These genes tend to show clear oscillations with high amplitude and with a period similar to the period of division of cell culture estimated from flow cytometry. The areas to the left and right consist of genes where the periodicity is less pronounced.

The genes in the upper part of the diagram show a peak expression in S and/or G2 phases, while the more numerous conserved cyclic genes clustered in the bottom showed peaks of expression in the end of G2, M and beginning of G1 phases.

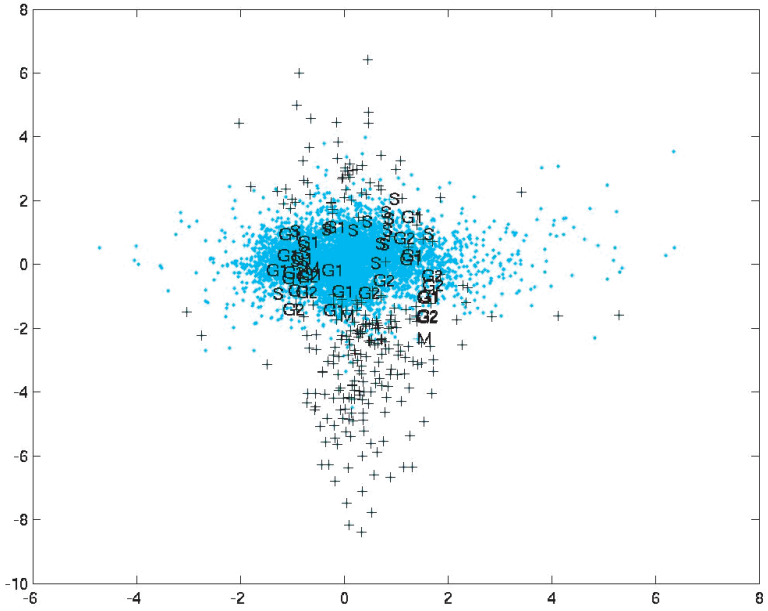

Figure 2: A visualization of genes significantly cyclic in human cells in the data of [19]. X and Y axes are two biggest principal coordinates (derived variables, which have the biggest influence on the overall data, see [4] for details). Genes with cyclic orthologs in other organisms are marked by black crosses.

3.1 Cyclic Groups of Orthologs

The groups of orthologs which are conserved among all genes include essential parts of cell cycle machinery. Conserved cyclic groups include cyclin B homologs, CDC5 kinase with multiple functions in mitosis and cytokinesis, SCH9 protein kinase regulating G1 progression, MCD1(SCC1) protein required for sister chromatin cohesion, DSK2 which is required for G2/M transit, ZPR1 cell cycle regulator and MIH1 phosphatase regulator of phosphorylation state of Cdc28p, transcription factor YHP1 - repressor at early cell cycle boxes and SDS22 subunit of serine-threonine phosphatases required for chromosome transmission.

The majority of proteins regulating the cell cycle of *S. cereviscae* belonged to conserved cyclic groups (Figure 3). However, there are also unconserved proteins. These include several transcription factors which are believed to be at the highest level of regulation, controlling cyclins and other proteins and also regulating each other in a ring of dependence [14]. Of these transcription factors, MCM1, SWI4, SWI6, MBP1, FKH1, FKH2 and NDD1 don't belong to conserved cyclic groups and SWI5 and ACE2 belong to a conserved cyclic group, but their closest homologs within the group are not cyclic.

Proteins acting in cytokinesis included myosin light chain and class II heavy chain, SMY1 kinesin, KIP2 kinesin-related motor protein, DBF20 Serine Threonine kinase and HOF1 protein required for cytokinesis. Components of chromatin include histones H2A, H2B, H3 and H4, and their regulators: NAP1 nucleosome assembly protein and SDS3 component of histone deacetylase complex. Genes regulating DNA damage and repair included Rad53 kinase required for cell-cycle arrest in response to DNA damage, MSH6 mismatch repair ATPase and rad51 DNA repair protein. Curiously, only some of checkpoint kinases and related serine/threonine proteins were conserved. Some of cyclic, conserved genes related to DNA replication were RFC2 replication factor and CDC45 DNA replication initiation factor. Cyclic were also BDF2 transcription initiation factor of TFIID and TFIID subunits 90 kDa

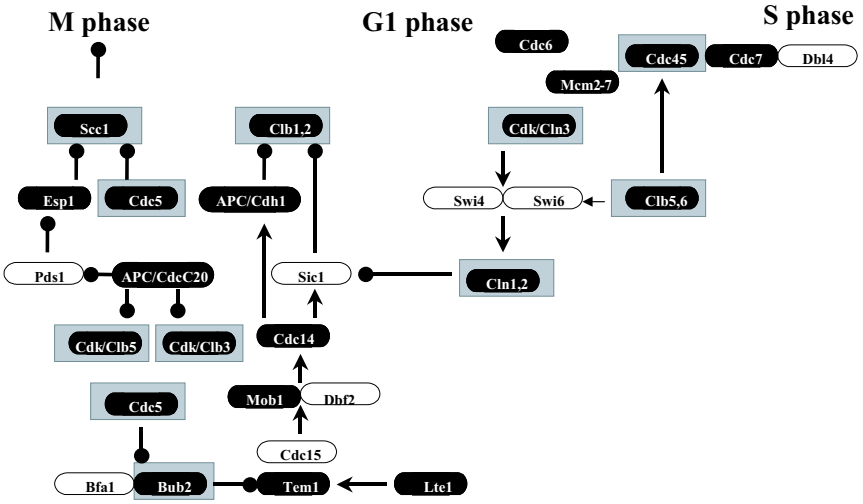

Figure 3: Genes regulating the cell cycle in *S. cereviscae,* adapted from [1]. Positive regulation is marked by arrows, negative by T-bars. Proteins with orthologs in *S. cereviscae*, *S. pombe* and *H. sapiens* genomes are marked by black ovals. Proteins without such sequence conservation are white ovals. Proteins belonging to conserved cyclic orthologs are shown on grey-shaded fields.

and RMTAF4.

Several conserved cyclic genes were related to translation, like genes of ribosomal proteins L10A, L5 and L11, L34, L7, L26, S27; translation initiation factors 3 and 4F, ARC1 t-RNA binding protein, DED1 Atp-dependent RNA helicase as well as numerous chaperones.Many proteins were involved in different processes related to the cell cycle, like UBC4 ubiquitin like protein and RSP5 ubiquitin protein ligase, which also has a role in chromatin assembly.

Several big protein families contain several cyclic and non-cyclic homologs. The large family of PP1 Serine threonine specific protein phosphatases contains two cyclic proteins in *S. cereviscae*, three in *S. pombe* and three in *H.sapiens*. Apparently, Two *S. cereviscae* proteins: cyclic ppz2 and uncyclic ppz1 are closely related to cyclic ppz in *S. pombe*. The second group of closely related proteins are PP11 protein in *S. pombe* and human protein phosphatase, together with not significantly cyclic glc7 protein in *S. cereviscae*. Several cyclic proteins lack very close homologs: ppq1 in *S. cereviscae* ppp1 in *H. sapiens* and O74480 in *S. pombe*.

Another group is NDR and related serine threonine kinases. It has cyclic and non-cyclic proteins. Cyclic stk38 protein in human forms a different subgroup from others. Cyclic *S. cereviscae* protein dbf20 involved in late nuclear division has non significantly cyclic homolog in *S. cereviscae* (dbf2) and in S. pombe sid2. Other, noncyclic proteins similar to NDR are in different subgroups.

In another group containing Chk2 Serine/threonine protein kinases, cyclic proteins Rad53 in *S. cereviscae*, CDS1 in *S.pombe* and CHK2 in *H.sapiens* are a different subgroup from another cyclic protein dun1 in the budding yeast.

Another big family of serine threonine protein kinases includes cyclic proteins. Here, two cyclic members in *S. cereviscae* (YPL141C and KCC4) as well as cyclic proteins in *S. pombe* and H.sapiens have closest similarity to non-cyclic kinases in other organisms. This suggests their largely independent function with respect to cell cycle.

Over 40 conserved cyclic groups of genes have functions not obviously related to the cell cycle. Examples of cyclic metabolic proteins are: SAR1 GTPase required for transport vesicle formation,

cytosolic sorting protein, peptidyl-prolyl cis-trans isomerase, polyprenyl synthetase; alkyl hydroperoxide reductase, CPR1 cyclophilin type peptidyl-prolyl isomerase, sterol reductase, fatty acid specific elongation enzyme. Presence of these cyclic proteins reflects dramatic changes in cell metabolism throughout cell cycle. Cyclic mitochondrial proteins included YHM4 DNA-binding protein; RML2 ribosomal protein,. RPO41 RNA polymerase and AAC1 ADP/ATP carrier proteins.

3.2 What is Conserved in Cell Cycle

The first comparison of cell cycle regulated proteins in budding and fission yeast [13] drew attention to the fact that cyclic genes showed little overlap between these. We confirmed that a set of conserved cyclic groups is limited. We conducted a simulation by drawing a number of genes at random from the genomes of three model organisms and counting shared orthologous groups. This test confirmed that such number of conserved cyclic orthologs is within expectation from the proportion of orthologs and cyclic genes in genomes.

Conserved cyclic groups contain proteins taking part in DNA replication, cell division, structural components of chromatin and DNA damage and repair. However, numerous conserved cyclic groups contain only proteins where function is apparently not related to cell cycle. This result is not surprising if we consider that cell metabolism changes greatly during the cell cycle. Some of these proteins take part in rebuilding cell components after division. Therefore, microarray and similar experiments should take into account that genes changing expression pattern in the cell cycle are not necessary taking part in cell cycle control.

Much of the cell cycle control proteins in *S. cereviscae* have orthologs in other model organisms (Figure 3). The conserved proteins include regulatory kinases, which diversified relatively recently during eukaryote evolution [6] and took over a role as important regulators of cell cycle. In contrast to this, many transcription factors are not conserved. In particular, this includes transcription factors regulating the cell cycle progression in *S. cereviscae* at the topmost level [14] which appears to be contradictory. One might speculate that the role of these transcription factors has been substituted by unrelated proteins, possibly exercising their effect rather through signaling than transcriptional regulation.

Cell cycle is a fundamental process in all organisms and therefore one might expect it to be extremely highly conserved. Therefore the lack of conservation of many important cell cycle proteins will remain interesting and hopefully provide clues as to the functioning of cellular regulation.

References

[1] Breeden, L.L., Periodic transcription: a cycle within a cycle, *Curr. Biol.*, 13:R31–38, 2003.

[2] 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., A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, 2:65–73, 1998.

[3] Cooper, S. and Shedden, K., Microarray analysis of gene expression during the cell cycle, *Cell Chromosome.*, 2:1, 2003.

[4] Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., and Vingron, M., Correspondence analysis applied to microarray data, *Proc. Natl. Acad. Sci. USA* , 98:10781–100786, 2001.

[5] Kishino, H. and Waddell, P.J., Correspondence analysis of genes and tissue types and finding genetic links from microarray data, *Genome Inform Ser Workshop Genome Inform*, 11:83–95, 2000.

[6] Krylov, D.M., Nasmyth, K., and Koonin, E.V., Evolution of eukaryotic cell cycle regulation: stepwise addition of regulatory kinases and late advent of the CDKs, *Curr. Biol.*, 13:173–177, 2003.

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

[8] Menges, M., Hennig, L., Gruissem, W., and Murray, J.A., Genome-wide gene expression in an Arabidopsis cell suspension, *Plant Mol. Biol.*, 53:423–442, 2003.

[9] Menges, M., De Jager, S.M., Gruissem, W., and Murray, J.A., Global analysis of the core cell cycle regulators of arabidopsis identifies novel genes, reveals multiple and highly specific profiles of expression and provides a coherent model for plant cell cycle control, *Plant J.* , 41:546–566, 2005.

[10] Ota, K., Goto, S., and Kanehisa, M., Comparative analysis of transcriptional regulation in eukaryotic cell cycles, *Fourth International Workshop on Bioinformatics and Systems Biology*, 2004.

[11] Peng, X., Karuturi, R.K., Miller, L.D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L.S., Liu, E.T., Balasubramanian, M.K, and Liu, J., Identification of cell cycle-regulated genes in fission yeast, *Mol. Biol. Cell*, 12:1026–1042, 2005.

[12] Remm, M., Storm, C.E.V., and Sonnhammer, E.L.L., Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *J. Mol. Biol.*, 314:1041–1052, 2001.

[13] Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bahler, J., Periodic gene expression program of the fission yeast cell cycle, *Nat. Genet.*, 36:809–817, 2004.

[14] 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., Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell*, 106:697–708, 2001.

[15] Sherlock, G., STARTing to recycle, *Nat. Genet.*, 36:795–796, 2004.

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

[17] Spellman, P.T. and Sherlock, G., Reply: whole-culture synchronization - effective tools for cell cycle studies, *Trends Biotechnol.*, 22:270–273, 2004.

[18] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., and Natale, D.A., The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, 4:41, 2003.

[19] 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., Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Mol. Biol. Cell*, 6:1977–2000, 2002.

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