Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, No. 7* **2823–2832**
doi:10.1093/nar/gkr1077

Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species

Miguel A. Fernández[1], Cristina Rueda[1] and Shyamal D. Peddada[2,*]

[1]Department of Statistics and Operations Research, Universidad de Valladolid, Prado de la Magdalena s.n., 47005 Valladolid, Spain and [2]Biostatistics Branch, National Institute of Environmental Health Sciences (NIEHS), Alexander Dr., RTP, NC 27709, USA

Received April 20, 2011; Revised September 28, 2011; Accepted October 28, 2011

ABSTRACT

A cell division cycle is a well-coordinated process in eukaryotes with cell cycle genes exhibiting a periodic expression over time. There is considerable interest among cell biologists to determine genes that are periodic in multiple organisms and whether such genes are also evolutionarily conserved in their relative order of time to peak expression. Interestingly, periodicity is not well-conserved evolutionarily. A conservative estimate of a number of periodic genes common to fission yeast (*Schizosaccharomyces pombe*) and budding yeast (*Saccharomyces cerevisiae*) ('core set FB') is 35, while those common to fission yeast and humans (*Homo sapiens*) ('core set FH') is 24. Using a novel statistical methodology, we discover that the relative order of peak expression is conserved in ~80% of FB genes and in ~40% of FH genes. We also discover that the order is evolutionarily conserved in six genes which are potentially the core set of signature cell cycle genes. These include *ace2* (a transcription factor) and polo-kinase *plo1*, which are well-known hubs of early M-phase clusters, *cdc18* a key component of pre-replication complexes, *mik1* which is critical for the establishment and maintenance of DNA damage check point, and histones *hhf1* and *hta2*.

INTRODUCTION

A cell division cycle among eukaryotes consists of a sequence of four major phases, namely, the G1, S, G2 and the M phase. The G1 phase (also known as the Gap 1 phase) is a resting phase where the cells grow in size and prepare for synthesis during the S phase. Furthermore, G1 phase also serves as one of two major check points, where if DNA damage is detected then the cell is prevented from proceeding to the S phase (1,2). Cells which pass the G1 check point proceed to S phase where the DNA replication takes place. This phase is followed by the G2 or Gap 2 phase. Similar to G1, this phase serves as a check point to ensure cells with damaged DNA do not proceed to the M phase (or mitosis) where the cells divide to form two daughter cells.

Genes participating in a cell division cycle have a cyclical pattern of expression with peak attained just before their function (3). There are several intrinsic differences among organisms in various aspects of cell division cycle. For instance, the amount of time spent by a cell in different phases varies. Fission yeast, *Schizosaccharomyces pombe* (*S. pombe*), cell spends almost 70% of its time in the G2 phase, whereas the budding yeast, *Saccharomyces cerevisiae* (*S. cerevisiae*), cell spends roughly quarter of its time in G2 phase. *Arbidopsis thaliana* has a relatively small G2 phase but a long S phase. In contrast to *S. pombe*, a human cell may spend substantially more time in the G1 phase than in G2 phase (See www.cyclebase.org). Also, the proportion of genes that are known to participate in cell division cycle varies with organisms. For example, it is estimated that there are twice as many periodic genes in *S. cerevisiae* as in *S. pombe* (4).

Despite many such differences, researchers are interested in (i) identifying genes that are periodic in multiple organisms (referred to as 'periodically conserved' genes) (Figure 1); (ii) among periodically conserved genes, identifying those that are also conserved in their phase of peak expression (Figure 1). This has been an active area of research over the past several decades [cf. (3,5)]. With the advent of microarray technology, numerous microarray studies have been conducted on several model organisms such as *S. cerevisiae* (6–9), *S. pombe* (4,10–12), *Homo sapiens* (13) and *Arabidopsis* (14,15). Such

*To whom correspondence should be addressed. Tel: +1 919 541 1122; Fax: +1 919 541 4311; Email: peddada@niehs.nih.gov

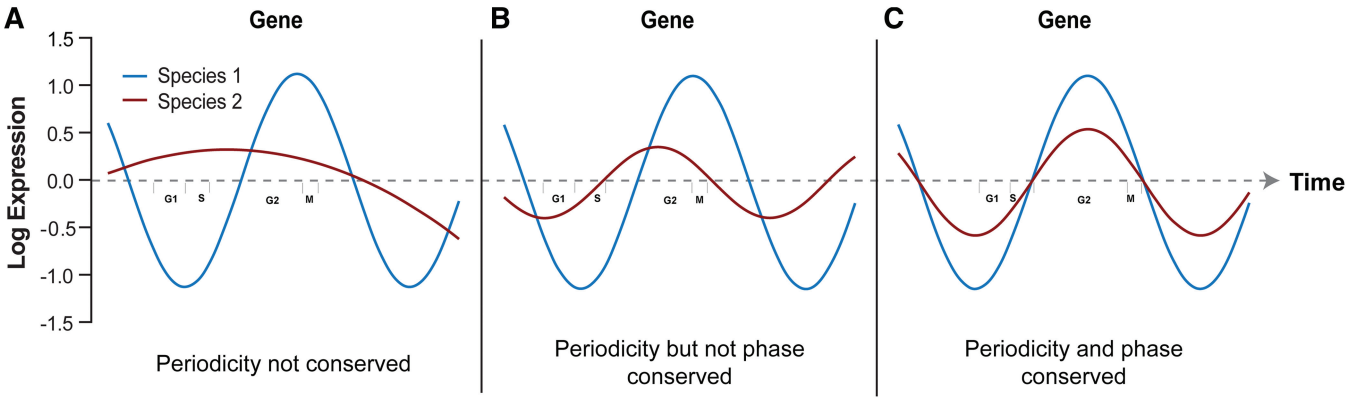

Figure 1. Conservation of periodicity and phase. In each panel, the vertical hash mark on the time axis represents the boundary of a phase.

large-scale genome wide data on multiple organisms provide an excellent opportunity to determine genes involved in cell cycle and study their functions. It also allows one to understand the similarities and dissimilarities in the cell cycle of various organisms. A useful database containing results from various cell cycle microarray experiments is available at www.cyclebase.org (16), henceforth referred as 'cyclebase'. These microarray data allow biologists to debate the conservation of genes participating in a cell cycle and their times of peak expression (3,4,10,11, 12,17). Based on a comprehensive analysis of the above microarray data and other published data, Jensen *et al.* (3) concluded that both periodicity as well as phase of peak expression are evolutionarily poorly conserved. Earlier a similar conclusion was drawn by Rustici *et al.* (4) who concluded that only 40 or so orthologs are periodic in both species of yeasts.

Although the poor conservation of periodicity and the phase of peak expression for most cell cycle genes may be biologically plausible, as evolutionarily functions of some genes may have changed, one cannot ignore variability between and within studies that may have contributed to these findings. For instance, even within the same organism there are major differences among studies published in the literature. Recently, three different groups of researchers conducted a total of 10 microarray experiments on *S. pombe* [5 by Rustici *et al.* (4), 3 by Oliva *et al.* (11) and 2 by Peng *et al.* (12)]. As summarized by (11) and by Caretta-Cartozo *et al.* (17), the three studies disagreed considerably on the number of periodic genes. According to Caretta-Cartozo *et al.* (17), only 156 out of ~5000 genes in *S. pombe* genome were declared to be periodic by all three studies, although individually each group identified at least three times as many periodic genes. Furthermore, even among genes that were found to be periodic in at least two of the three studies, there were disagreements among the studies in terms of time to peak expression of some genes. For instance, according to Peng *et al.* (12) *cdc18*, *mob1*, *imp2* and *cig2* peak during the G1 phase, whereas Oliva *et al.* (11), Rustici *et al.* (4) and cyclebase suggest that these are M phase genes.

The above discrepancies among studies even in the same organism are not surprising and can be attributed to various factors such as, natural variability in the data, experimental conditions, etc. Factors such as these result in statistical variability and uncertainty in estimates of time to peak expression. Not much has been discussed in the literature regarding these issues. The problem becomes even more challenging when comparisons across multiple organisms are to be made. In such comparisons, the biological differences among organisms may be confounded by statistical uncertainties due to variability in the data. For example, Rustici *et al.* (4) concluded that cell cycle regulation of majority of genes is not conserved between *S. pombe* and *S. cerevisiae*. On the other hand, Oliva *et al.* (11) and Peng *et al.* (12) suggest greater amount of similarities between the two species of yeasts and infer conservation of regulatory mechanisms.

Since cell division cycle is a carefully orchestrated process, it is reasonable to hypothesize that the relative order of peak expression among a core set of cell cycle genes may be conserved even if the phase of some genes may have been evolutionarily modified (Figure 2). Genes whose relative order of time to peak expression is conserved may not only have a well-defined function in cell division cycle, but may also have potential interactions or associations with each other. In this article, we develop a formal statistical methodology to test the hypothesis that the relative order of peak expression among a core set of cell cycle genes is conserved in a pair of organisms. Using this methodology, we investigate if the 'core set FB' of fission yeast genes have the same relative order of peak expression as their budding yeast orthologs. Similarly, we investigate if the 'core set FH' of fission yeast genes have the same relative order of peak expression as their human orthologs. The statistical procedure developed in this article is novel and could help biologists formulate and test other similar hypotheses.

MATERIALS AND METHODS

Relative order of peak expression among cell cycle genes

For a gene D, the time to its peak expression can be described in terms of an angle on a unit circle, known as the phase angle, which is denoted by ϕ_D. Suppose D, E and F are three cell cycle genes where D is an S

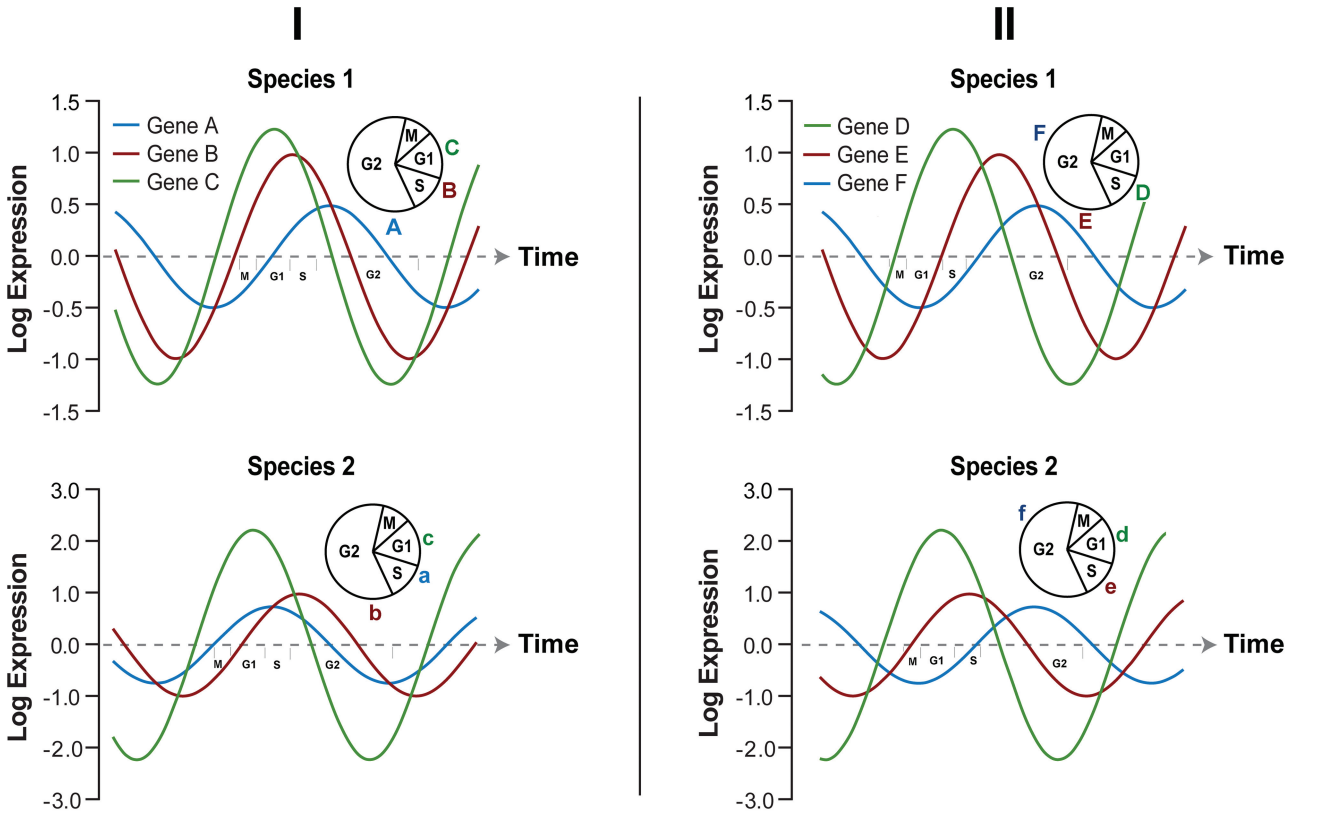

Figure 2. (**I**) Relative order of peak expression of genes A,B and C is not conserved in Species 1 and 2. (**II**) Relative order of peak expression of genes D, E and F is conserved in Species 1 and 2. Orthologs in Species 2 are denoted by lower case letters a, b, c, d, e and f. In each panel, the vertical hash mark on the time axis represents the boundary of a phase. Insets in each panel represent the time to peak expression in terms of phase of cell cycle.

phase gene, E is an early G2 phase gene and F is a mid-G2 phase gene, then they satisfy the relative order; D followed by E which is followed by F and which is followed by D. We represent this relative order of peak expression among the three genes by $\phi_D \prec \phi_E \prec \phi_F \prec \phi_D$ (or equivalently, $D \prec E \prec F \prec D$). Suppose *S. cerevisiae* genes D, E and F are the orthologs of *S. pombe* genes d, e and f, respectively. Suppose $D \prec E \prec F \prec D$, then we say that the relative order is conserved among the orthologs if $d \prec e \prec f \prec d$ (see Figure 2). In the ideal setting, if functions of all genes in the core set are conserved through evolution and if cell cycle is a well-ordered mechanism of nature, then it is reasonable to hypothesize that the relative order of expression of genes in the core set is conserved between the two organisms. Note that the relative order is invariant of the location of the pole of the circle. This is important for several reasons. First, biologically, the order of genes around the circle has no bearing on where the pole of the circle is established (i.e. it is rotation invariant). Secondly, a common challenge with time course cell cycle experiments is that one cannot be sure about the exact biological time when the cells were arrested to define the pole precisely. Also, different labs and experiments may arrest cells during different phases of cell cycle. Consequently, it can be challenging to compare phase angles across experiments since each experiment may have a different pole. However, the relative order of genes should be invariant of the location of the pole. Also, it is important to note that our definition of conservation of relative order does not require the orthologs pairs (D, d), (E, e), (F, f), etc. to have same phase angles or even the same phases (see the right panels in Figure 2). We just require d, e, f to satisfy the same relative order as D, E, F.

Using Rustici *et al.* (4), Oliva *et al.* (11) and cyclebase, we arrived at the core set FB of 35 *S. pombe* cell cycle genes that are periodic in both yeasts (Table 1). Similarly, using cyclebase we arrived at the core set FH of 24 *S. pombe* cell cycle genes that are periodic in both *S. pombe* as well as *H. sapiens* (Table 2). We limited our core sets to include only those genes whose cycelbase periodicity rank is $\leq$500. The rank cut-off of 500 was arbitrarily chosen. Our point is that genes with higher ranks are less likely to be periodic with estimated phase angles subject to small concentration parameter, resulting in large uncertainty estimates.

In the case of human orthologs, we relied completely on the peaktime specified by the cyclebase database to arrive their relative order (Table 2). However, in the case of budding yeast orthologs, in addition to cyclebase, we used published literature and Saccharomyces Genome Database (http://www.yeastgenome.org/) to arrive at the relative order (Table 1).

We now describe the relative order of the 35 *S. cerevisiae* orthologs. Since the mRNA level as well

Table 1. *Saccharomyces pombe* genes in the core set FB arranged according to the relative order of *S. cerevisiae* orthologs

Relative order	*Saccharomyces cerevisiae*				*Saccharomyces pombe*		
	Gene	CBase Rank	Peaktime (CBase) (deg.)	Phase (CBase)	Gene	CBase Rank	FSA order
1	CDC6	500	349.2	M	*cdc18*	12	18
2	RFA1	6	68.4	G1	*ssb1*	56	12
3	RNR1	54	72	G1	*cdc22*	8	2
4	MSH6	16	72	G1	*msh6*	50	10
5	MRC1	192	72	G1	*mrc1*	33	23
6	POL1	112	72	G1	*pol1*	61	27
7	SMC3	55	72	G1	*psm3*	73	14
8	MCD1	11	75.6	G1	*rad21*	94	16
9	RAD51	34	75.6	G1	*rhp51*	275	34
10	CLN2	8	79.2	G1	*cig2*	38	9
11	POL2	84	79.2	G1	*pol2*	128	32
12	SWE1	77	97.2	S	*mik1*	51	11
13	HHT2	28	133.2	S	*h3.3*	20	5
14	HHF1	30	136.8	S	*hhf1*	17	3
15	HHT1	13	140.4	S	*hht3*	26	6
16	HTA2	12	144	S	*hta2*	31	7
17	HTB2	5	144	S	*htb1*	19	4
18	HTZ1	188	176.4	S	*pht1*	122	31
19	KIP3	390	165.6	S	*klp5*	69	28
20	FKH1	85	180	S/G2	*fkh2*	35	8
21	SWI5	124	234	G2	*ace2*	15	19
22	BUD4	229	234	G2	*mid2*	28	22
23	CDC5	117	234	G2	*plo1*	46	24
24	MOB1	177	248.4	G2	*mob1*	119	30
25	ASE1	111	252	G2	*mcp1*	359	35
26	MYO1	253	252	G2	*myo3*	49	26
27	CHS2	87	252	G2	*chs2*	59	13
28	HOF1	113	255.6	G2	*cdc15*	11	17
29	HOF1	113	255.6	G2	*imp2*	76	29
30	KIN3	104	291.6	G2/M	*fin1*	47	25
31	DBF2	72	298.8	M	*sid2*	83	15
32	SST2	471	324	M	*rgs1*	159	33
33	CDC20	24	338.4	M	*slp1*	2	1
34	PST1	100	0	M/G1	*SPAC1705.03C*	24	21
35	DSE4	435	21.6	G1	*eng1*	21	20

CBase = Cyclebase

as its protein level peaks during the early stages of G1 phase and is the precursor for DNA synthesis, therefore we begin with CDC6. This gene is followed by several G1 phase genes such as those involved in DNA repair, replication and check point (Replication Factor Alpha, RNR1, MSH6, MRC1 and POL1), cohesion of sister chromatids (SMC3 and MCD1), recombinational repair of double-strand breaks in DNA (RAD51), activation of Cdc28p to promote transition from G1 to S phase (CLN2, a late G1 phase cyclin) and DNA synthesis during DNA repair (POL2). The S phase genes that followed the G1 phase genes are those involved in: regulation of G2/M transition by inhibition of Cdc28p kinase activity (SWE1), chromatin assembly (histones such as HHT2, HHF1, HHT1, HTA2, HTB2, HTZ1) and mitotic spindle position (KIP3). These are followed by G2 phase transcription factors such as FKH1 and SWI5. Several G2 phase genes considered here have proteins involved in important functions such as: bud site selection (BUD4), cytokinesis and septation (CDC5, MOB1, ASE1, MYO1, CHS2, HOF1). Note that we considered both *S. pombe* orthologs of HOF1, namely, *cdc15* and *imp2* in our analysis. These genes are followed by KIN3, G2/M check point gene whose protein Kin3 plays a critical role in DNA damage recognition before the cell enters M phase (18). Among the M phase genes in the proposed relative order, DBF2 and CDC20 have a function for cells to exit from mitosis, while PST1 has function in the construction of cell wall. Our proposed relative order concluded with DSE4, a daughter cell-specific protein which degrades the cell wall causing the daughter cell to separate from the mother cell. Hence, it is logical that DSE4 was the last gene in our proposed relative order before returning to CDC6. According to cyclebase, some of the *S. cerevisiae* orthologs in the core set FB have identical peaktimes hence we assigned identical phase angles to all such genes in the null hypothesis. Specifically, following genes within parenthesis were hypothesized to have the same phase angles: (*cdc22*, *msh6*, *mrc1*, *pol1*, *psm3*), (*rad21*, *rhp51*), (*cig2*, *pol2*), (*ace2*, *mid2*, *plo1*), (*mcp1*, *myo3*, *chs2*) and (*cdc15*, *imp2*). The resulting relative order is described in Figure 3, where genes that were hypothesized to have same phase angle are along the same ray from the center of the circle (ray not drawn). See also Table 1. Known biological functions of genes in the core set FB are provided

Table 2. *Saccharomyces pombe* genes arranged in the core set FH according to the relative order of *H. sapiens* orthologs

Relative order	*Homo sapiens*				*Saccharomyces pombe*		
	Gene	CBase Rank	Peaktime (CBase) (deg.)	Phase (CBase)	Gene	CBase rank	FSA order
1	IFIT2	313	10.8	G1	*ssn6*	249	19
2	LRRC56	496	43.2	G1	*sds22*	384	22
3	ZNF367	48	140.4	G1	*ace2**	15	3
4	CDC6	40	165.6	G1	*cdc18**	12	2
5	PKMYT1	231	187.2	S	*mik1**	51	13
6	HIST2H4B	219	194.4	S	*hhf1**	17	11
7	DHFRL1	213	208.8	S	*dfr1*	473	24
8	SH3GL2	332	219.6	S	*SPBC19C2.10*	310	21
9	H2AFX	83	259.2	S/G2	*hta2**	31	4
10	HRSP12	116	262.8	G2	*mug71*	155	14
11	PCBP1	189	266.4	G2	*rnc1*	186	16
12	TOP2A	14	291.6	G2	*top2*	226	18
13	PIF1	35	302.4	G2	*pif1*	88	8
14	FOXM1	107	309.6	G2	*fkh2**	35	12
15	Hsp40	203	320.4	G2	*SPCC63.13*	216	17
16	CDC25B	64	324	G2/M	*cdc25*	158	15
17	AURKA	4	324	G2/M	*ark1*	265	20
18	KIF10	7	327.6	M	*klp5**	69	7
19	CCNB1	49	334.8	M	*cig2**	38	5
20	PLK1	1	334.8	M	*plo1**	46	6
21	MAPK13	3	334.8	M	*spm1*	97	10
22	CDC20	24	334.8	G2	*slp1**	2	1
23	API4	38	338.4	M	*bir1*	453	23
24	RAD21	80	345.6	M	*rad21**	94	9

Genes with * have periodic *S. cerevisiae* orthologs, CBase = Cyclebase

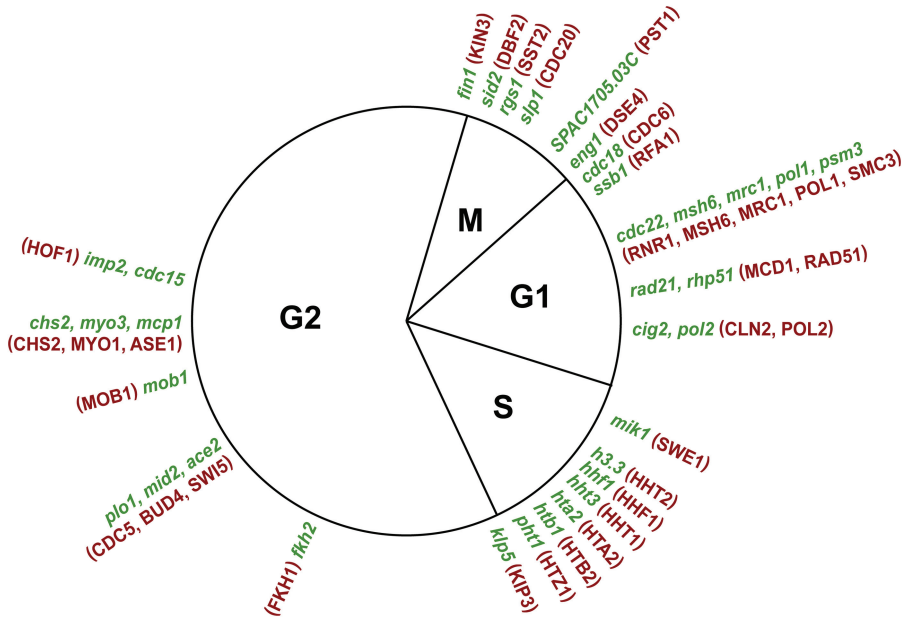

Figure 3. *Saccharomyces pombe* genes arranged according to the relative order and approximate locations of their *S. cerevisiae* orthologs (in parenthesis). Sectors drawn are according to *S. pombe* cell cycle.

in Supplementary Table S1. The goal of this study is to test whether the *S. pombe* genes satisfy the relative order specified by the *S. cerevisiae* orthologs. Thus we tested the null hypothesis that the phase angle of *cdc18* is followed by the phase angle of *ssb1*, ..., *SPAC1705.03c* which in turn is followed by the phase angle of *eng1* which in turn is followed by the phase angle of *cdc18* against the alternative hypothesis that this order is not true. More precisely:

$$H_0 : \phi_{cdc18} \prec \phi_{ssb1} \prec \phi_{cdc22} = \phi_{msh6} = \ldots = \phi_{psm3} \prec \ldots \prec \phi_{SPAC1705.03c} \prec \phi_{eng1} \prec \phi_{cdc18} \quad (1)$$

$$H_1 : H_0 \text{ is not true}$$

Similarly, we tested the following hypotheses to see whether the *S. pombe* genes satisfy the relative order specified by the 24 *H. sapiens* orthologs:

$$H_0 : \phi_{ssn6} \prec \phi_{sds22} \prec \phi_{ace2} \prec \ldots \prec \phi_{plo1} = \ldots = \phi_{cig2} \prec \phi_{bir1} = \phi_{slp1} \prec \phi_{rad21} \prec \phi_{ssn6} \quad (2)$$
$$H_1 : H_0 \text{ is not true}$$

There are two reasons (biological and statistical) for the above formulation of null and alternative hypotheses. First, as stated earlier, the cell division cycle is a fundamental process in eukaryotes and one would expect various aspects of this process to be conserved through evolution. This is the basic premise of many recent papers [e.g. (4)] which tried to identify genes that are periodic in multiple organisms. Since our investigation is based on such conserved genes, the conservation of the relative order should be the null hypothesis rather than the alternative hypothesis. Basically, among genes that are declared to be conserved between species, our null hypothesis states that their relative order is also conserved. There is also a statistical reason for our choice of null and alternative hypotheses. If the null hypothesis was that the relative order is not conserved among the q genes in the two organisms, then the null hypotheses would contain $(q-1)!$ configurations of parameters and the alternative hypothesis contains only one configuration. As q increases, the null hypothesis is too large and is never likely to be rejected. For example, if $q = 35$ the total number of possible null configurations are of the order 10^{40}, which is extremely large. No statistical test would have sufficient power to reject the null hypothesis in such situations. In fact, the power will go to zero as q increases!

Statistical test

For each gene i, $i = 1, 2, \ldots, q$ and experiment j, $j = 1, 2, \ldots, E$, we model the unconstrained estimator of phase angle of peak expression θ_{ij}, obtained from the Random Period Model (RPM) (19), using the *von Mises* distribution (*VM*). This distribution plays an important role in circular data analysis similar to the normal distribution for Euclidean data. Thus, we assume that $\theta_{ij} \rightsquigarrow VM(\phi_{ij}, \kappa_j)$ where ϕ_{ij} is the true unknown phase angle of peak expression of gene i in the j-th experiment. The concentration parameter κ_j represents the uncertainty associated with θ_{ij}. We assume that κ_j depends on experiment j but not on gene i. There are two sources of uncertainty associated with the phase angle estimate of each gene, one is specific to the gene and the other is due to the experiment (which is common to all genes within the experiment). This resembles the classical mixed effects linear model in Euclidean space data. Since the number of time points used in each of the time course experiments considered in this article is fairly large, for any specific gene, the uncertainty associated with the estimator of the phase angle based on the RPM is negligible relative to the uncertainty due to the experiment. For this reason, we only retained the uncertainty component corresponding to the experiment.

For each experiment j, $j = 1, 2, \ldots, E$, our problem of interest is to the test the following hypotheses:

$$H_0 : \text{The phase angles } \phi_{ij},\ i = 1, \ldots, q \text{ follow a known relative order} \quad (3)$$
$$H_1 : H_0 \text{ is not true.}$$

Let $C_q^I = \{x \in \Re^q : 0 \le x_I \le .. \le x_q \le x_1 \le .. \le x_{I-1} \le 2\pi\}$ be a simple order constraint starting at index I and let $\phi_j = (\phi_{1j}, \phi_{2j}, \ldots, \phi_{qj})'$. Then, for each $j = 1, 2, \ldots, E$, the above null hypothesis can be rewritten as $\phi_j \in C_q$, where $C_q = \bigcup_{1 \le I \le q} C_q^I$.

Let $\widetilde{\theta}_j = (\widetilde{\theta}_{1j}, \widetilde{\theta}_{2j}, \ldots, \widetilde{\theta}_{qj})'$ denote the restricted maximum likelihood estimator of ϕ_{ij} subject to the constraint $\phi_j \in C_q$ (20). $\widetilde{\theta}_j$ determines a partition $f = \{1, \ldots, I\}$ into sets of consecutive coordinates on which $\widetilde{\theta}_j$ is constant. These sets are called level sets. Then, we construct the following test statistic to test the above hypotheses

$$T_j = 2\kappa_j \sum_{i=1}^{q} \left(1 - \cos\left(\theta_{ij} - \widetilde{\theta}_{ij}\right)\right).$$

Notice that T_j is a measure of the angular distance between θ_j and $\widetilde{\theta}_j$. Our conditional test T_j, described in detail in the Supplementary Data, rejects the null hypothesis whenever $T_j \ge c(m)$, where m is the number of level sets for $\widetilde{\theta}_j$, $c(m)$ is chosen so that $pr(\chi^2_{q-m} \ge c(m)) = \alpha/(1 - pr_{\phi^0}(C_q))$, and ϕ^0 is any point in the null hypotheses for which all coordinates are equal. Since in practice, κ_j are unknown, in order to derive a proper value for these parameters, we obtained its maximum likelihood estimator using an analysis of variance approach based on the *von Mises* distribution (see Section 1.3 in Supplementary Data).

Theoretical details of T_j are given in Section 1.4 of Supplementary Data. There we demonstrate that our proposed test is an asymptotic α level test. The derivation of this latter property is not straightforward as several statistical issues arise as a result of the specific characteristic of the testing problem, namely, a complicated null hypothesis for a directional parametric model.

Lack of fit criterion for a given relative order

For a relative order specified by the null hypothesis, let p_j denote the P-value associated with the j-th experiment, $j = 1, 2, \ldots, E$, and let $L = -\sum_{j=1}^{E} \log(p_j)$. Note that L always lies between 0 and ∞, with smaller value corresponding to better fit. In the extreme case, if the presumed relative order is perfectly satisfied within each experiment with P-value of 1, then $L = 0$. Thus, among a collection of plausible orders for a set of cell cycle genes, a biologist may choose the order that corresponds to the smallest value of L. Note that under the null hypothesis, if the P-values are independently and uniformly distributed in the interval (0, 1), then $2L$ is distributed as a central χ^2 random variable with E degrees of freedom. This is often known as Fisher's method of combining P-values and yields a formal statistical test.

RESULTS

Using the 10 *S. pombe* time course experimental data (4,11,12), we first obtained the unconstrained phase angle estimates of genes in the core sets FB and FH (Supplementary Tables S2 and S3) which are then used for testing various hypotheses described in this article. Using the estimates in Supplementary Table S2, we tested the hypotheses appearing in Equation (1) that all 35 *S. pombe* genes in FB satisfy the relative order specified by the *S. cerevisiae* orthologs against the alternative hypothesis that they are not. The null hypothesis is rejected at $P \leq 0.15$ in 5 out of 10 experiments (Table 3). Of these five experiments, two have a $P < 0.0001$. If the null hypothesis was true in each of the 10 experiments, then the binomial probability of observing two or more experiments (out of 10) with a $P = 0.0001$ is 4.49×10^{-7}, which is extremely small. This suggests that the relative order hypothesized in Equation (1) may not be true and thus the 35 *S. pombe* genes do not follow the same relative order as their *S. cerevisiae* orthologs. Of course, in the above argument we implicitly assume that the outcomes of the 10 experiments are identically and independently distributed. Although this is a commonly made assumption, we acknowledge that it may be restrictive.

A question of interest is whether we can identify a subset of the 35 genes that conserve the relative order between the two yeasts. Since the number of all possible subsets (of various sizes) is extremely large, it would be practically impossible to enumerate all possible subsets of all sizes and then test the null hypotheses such as the one appearing in Equation (1) for each subset. This problem resembles the classical problem of selection of variables (or model selection) in linear regression analysis. Accordingly, we developed a Forward Selection Algorithm (FSA), which is described in the Supplementary Data. Similar to forward selection procedure in classical linear regression analysis, the FSA proceeds systematically by entering one gene at a time into the test for relative order according to its periodicity rank assigned by the cyclebase. Smaller the rank, the more periodic the gene is and hence its phase angle estimate is more likely to be reliable. The proposed FSA begins with all ortholog pairs that have a cyclebase rank <100. Thus, a gene is included in Step 1 of FSA if both fission yeast as well as the budding yeast orthologs of the gene has a rank <100. Details of the subsequent steps and the implementation of FSA are provided in the Supplementary Data.

Using FSA (Supplementary Table S4), we discover that 28 out of 35 *S. pombe* genes, namely, *cdc18*, *ssb1*, *cdc22*, *msh6*, *mrc1*, *pol1*, *psm3*, *rad21*, *cig2*, *pol2*, *mik1*, *h3.3*, *hhf1*, *hht3*, *hta2*, *htb1*, *pht1*, *klp5*, *fkh2*, *ace2*, *plo1*, *chs2*, *cdc15*, *imp2*, *sid2*, *slp1*, *SPAC1705.03C*, *eng1*, potentially satisfy the same order as their *S. cerevisiae* orthologs. Thus, the relative order of these 28 genes seems to be conserved between the two species of yeasts. For these genes, the null hypothesis is rejected in none of the experiments even at a level of significance as high as 0.30 (Table 3). It is also interesting to note that the lack of fit criterion L based on all 35 genes was 46.75 and it dropped to 3.57 for the above 28 genes selected by FSA.

Similar to genes in FB, we also tested the Equation (2) for genes in the core set FH and found that the relative order was rejected in 6 out of 10 experiments at a $P < 0.001$ (Table 4). Using FSA we found *ace2*, *cdc18*, *mik1*, *histones* (*hhf1*, *hta2*), *rnc1*, *top2*, *cdc25*, *plo1* and *slp1*, to satisfy the same relative order as their human orthologs (Table 4). Among these 10 genes, *ace2*, *cdc18*, *mik1*, *histones* (*hhf1*, *hta2*) and *plo1* also satisfied the relative order specified by their *S. cerevisiae* orthologs. Recall that, evolutionarily, humans and fungi are ~1.5 billion years apart and budding yeast and fission yeasts are nearly billion years apart (21). Thus, it appears that the above six genes are evolutionarily conserved in their relative order of peak expression during the cell division cycle (Figure 4 and Table 5). These six genes are well known in the literature to play a critical role during cell division cycle. For example, the transcription factor *a*ce2 and the polo-kinase *plo1* are well-known hubs of early M phase clusters (22), the cell cycle gene *cdc18* is a key component of pre-replication complexes for the onset of S phase (23), histones *hhf1*, *hta2* play an important role

Table 3. Test for relative order of *S. pombe* genes in the core set FB (Order specified by *S. cerevisiae* orthologs)

Experiment	*P*-values	
	Based on all 35 *Saccharomyces pombe* genes	Based on final 28 *Saccharomyces pombe* genes
Oliva cdc	0.08	0.53
Oliva elut1	0.69	0.98
Oliva elut2	0.14	0.41
Peng cdc	0.24	0.99
Peng elut	0.57	0.99
Rust cdc1	5.37E-11	0.34
Rust cdc2	0.19	0.86
Rust elut1	0.07	0.88
Rust elut2	0.24	0.99
Rust elut3	2.88E-05	0.53
Lack of fit	46.75	3.57

Table 4. Test for relative order of *S. pombe* genes in the core set FH in the 10 experiments (Order specified by *H. sapiens* orthologs)

Experiment	*P*-values	
	Based on all 24 *Saccharomyces pombe* genes	Based on final 10 *Saccharomyces pombe* genes
Oliva cdc	0.03	0.60
Oliva elut1	0.10	0.93
Oliva elut2	0.19	0.95
Peng cdc	1.10E-03	0.95
Peng elut	0.07	0.83
Rust cdc1	4.12E-10	1
Rust cdc2	2.37E-06	0.93
Rust elut1	0	0.92
Rust elut2	1.90E-13	0.93
Rust elut3	0.06	1
Lack of fit	>100 000	1.10

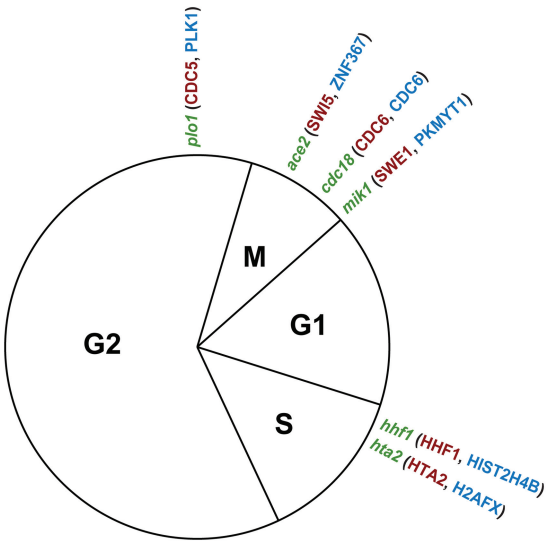

Figure 4. A core set of signature cell cycle genes with relative order of time to peak expression conserved among *S. pombe*, *S. cerevisae* and *H. sapiens*. Sectors and approximate locations of genes are drawn according to *S. pombe*. *S. pombe* genes are in green, *S. cerevisiae* orthologs are in red and *H. sapiens* orthologs are in blue.

Table 5. A core set of signature cell cycle genes

Saccharomyces pombe gene (*Saccharomyces cerevisiae*, *Homo sapiens* orthologs)	*Saccharomyces pombe*	*Saccharomyces cerevisiae*	*Homo sapiens*
plo1 (CDC5, PLK1)	G2	G2	M
ace2 (SWI5, ZNF367)	G2/M	G2	G1
cdc18 (CDC6, CDC6)	M	M	G1/S
mik1 (SWE1, PKMYT1)	M	G1/S	S
hhf1 (HHF1, HIST2H4B)	G1/S	S	S
hta2 (HTA2, H2AFX)	G1/S	S	S/G2

Cell cycle phases are obtained from cyclebase

during the S phase and *mik1* is critical in the establishment and maintenance of DNA damage check point (24).

To ensure that our statistical test has sufficient power to detect the alternative hypothesis, i.e. reject the null hypothesis that the genes in both species satisfy the same relative order, we conducted a simulation study for the fission and budding yeast data by randomly permuting the order of the genes in Step 1 of FSA and applied the algorithm. We considered 100 permutations and performed the first step of FSA on each permuted data. The null was rejected for all 100 permutations. We also found that in at least 5 out of the 10 experiments the P <0.05 and this occurred in every one of the 100 random permutations we considered. Note that the binomial probability of observing a *P*-value of 0.05 in at least 5 experiments out of 10 experiments by random chance is 6.36×10^{-5}, which is a very unlikely event. Yet, in all 100 random permutations we found 5 out of 10 experiments to have a $P < 0.05$, thus suggesting that our test is reasonably powerful to reject the null hypothesis of relative order if the hypothesis is not true. In our simulation study, we did not investigate the power of our test for alternatives where the order among the genes is not well conserved but not entirely random order. As with any statistical test, there will be a reduction in power as we get closer to the null hypothesis. In other words, if the true order is a very minor perturbation of the null hypothesis then probability of rejecting the null hypothesis would be smaller than when true order is substantially different from the null hypothesis. In a future project, we plan to investigate this problem in greater detail.

DISCUSSION

Since cell cycle genes follow a synchronized pattern of expression (25), one may speculate that some of the cell cycle genes are functionally conserved through evolution. There is an intrinsic order to the peak expression among the cell cycle genes so that they are converted into proteins in a well-synchronized manner to execute their respective functions during cell cycle. Consequently, the relative timing of peak expression of some of the cell cycle genes must be conserved through evolution.

Often the order among genes is determined using heat maps and published literature. There does not seem to exist a formal statistical methodology to test hypothesis regarding the order among genes in a given experiment. In this article, we have developed a novel statistical methodology that can be used for testing relative order among the phase angles of cell cycle genes. Using the methodology developed in this article, we demonstrated that a core subset of 28 *S. pombe* genes have the same relative timing of peak expression as their *S. cerevisiae* orthologs. This number increases to 32 if we reduce the stringency of our criterion. Thus, it may be reasonable to infer that among the 35 genes in the core set FB, at least 80% satisfy the same relative order of peak expression as their *S. cerevisiae* orthologs. Similarly, ~40% of the FH core set genes (10 out of 24) satisfy the same relative order of peak expression as their *H. sapiens* orthologs.

Although this article takes the first step toward a formal statistical methodology for answering questions about conservation of the relative order cell cycle genes, it is important to acknowledge that, analogous to classical linear regression analysis, one may consider other alternatives to FSA and derive an improved algorithm.

In this article, we took a 'conservative approach' when formulating the hypotheses appearing in Equations (1) and (2). Note that the average cyclebase ranks of *S. cerevisiae* and *H. sapiens* orthologs used in this article are 126.57 and 121.21, respectively. These are almost twice the average cyclebase rank of *S. pombe* genes, which is 66.17. Since higher cyclebase rank corresponds to poorer periodicity, therefore potentially, there is greater uncertainty in the phase angles of *S. cerevisiae* and *H. sapiens* orthologs in comparison to *S. pombe* genes. Since we formulated our null hypotheses using *S. cerevisiae* and *H. sapiens* orthologs and used *S. pombe* data to test, the FSA is more likely to select fewer genes than otherwise. The above formulation resembles the classical 'single sample' statistical hypothesis testing problem. It would be useful to extend our procedure so that

uncertainties in both orthologs are taken into account when formulating the testing problem, resembling the classical 'two sample' testing problem. Note that the 'two-sample' problem for testing the equality of two sets of orderings is not well developed even for Euclidean space data, and the problem is substantially more complicated for circular data.

The proposed relative order for *S. cerevisiae* and *H. sapiens* were determined using the peak times reported in cyclebase and the published literature. We recognize that the exact order among some of the 'neighboring' genes is difficult to ascertain. Thus, there is a potential for misspecification of the relative order. This resembles the classical problem of model misspecification that occurs so commonly in a variety of situations. If a biologist chooses to refine our proposed relative order based on her/his understanding of the functions and order of the genes, then she/he may explore such alternative orders and test them using our proposed methodology. A biologist could also select best fitting relative orders using the lack of fit criterion introduced in this article. Hence in this article we have provided a general methodology that would allow biologists to hypothesize a sequential order of peak expression for cell cycle genes and test it.

A freely downloadable SAS based user-friendly software can be obtained by either contacting the first author or by visiting www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm. An R package containing this and other circular data analysis routines is being developed and will soon be made available at the above address.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. Supplementary Tables S1–S4. Supplementary Information and Supplementary References [26–31].

ACKNOWLEDGEMENTS

The authors thank Drs S. Pyne (Harvard University), S. Surapureddi (NIEHS) and R. Jothi (NIEHS) for their valuable comments on an earlier draft which have led to improved presentation of the manuscript. The authors also thank Ms Sue Edelstein (NIEHS) for preparing the figures provided in this article. We are greatful to the developers of the website cyclebase.org which was extremely useful in this research.

FUNDING

Spanish Ministerio de Ciencia e Innovación (MTM2009-11161 to M.A.F. and C.R.); Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (Z01 ES101744-04). Funding for open access charge: Biostatistics Branch, NIEHS, National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Hartwell,L. and Kastan,M. (1994) Cell cycle control and cancer. *Science*, **266**, 1821–1828.
2. Elledge,S. (1996) Cell cycle check points: preventing an identity crisis. *Science*, **274**, 1664–1671.
3. Jensen,L.J., Jensen,T.S., de Lichtenberg,U., Brunak,S. and Bork,P. (2006) Co-evolution of transcriptional and posttranslational cell-cycle regulation. *Nature*, **443**, 594–597.
4. Rustici,G., Mata,J., Kivinen,K., Lió,P., Penkett,C.J., Burns,G., Hayles,J., Brazma,A., Nurse,P. and Bähler,J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.
5. Bahler,J. (2005) Cell-cycle control of gene expression in budding and fission yeast. *Annu. Rev. Genet.*, **39**, 69–94.
6. Cho,R., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
7. Spellman,P., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
8. de Lichtenberg,U., Wernersson,R., Jensen,T.S., Nielsen,H.B., Fausbøll,A., Schmidt,P., Hansen,F.B., Knudsen,S. and Brunak,S. (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, **22**, 1191–1201.
9. Pramila,T., Wu,W., Miles,S., Noble,W.S. and Breeden,L.L. (2006) The forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.*, **20**, 2266–2278.
10. Rustici,G. (2004) Periodic gene expression program of the fission yeast cell cycle. *PhD Dissertation*. Darwin College, Cambridge University, Cambridge, UK.
11. Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B. and Leatherwood,J. (2005) The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.*, **3**, 1239–1260.
12. Peng,X., Karuturi,R.K.M., Miller,L.D., Lin,K., Jia,Y., Kondu,P., Wang,L., Wong,L.S., Liu,E.T., Balasubramanian,M.K. *et al.* (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.
13. Whitfield,M., 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. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
14. Menges,M., Hennig,L., Gruissem,W. and Murray,J.A. (2002) Cell cycle-regulated gene expression in Arabidopsis. *J. Biol. Chem.*, **277**, 41987–42002.
15. Menges,M., Hennig,L., Gruissem,W. and Murray,J.A. (2003) Genome-wide gene expression in an Arabidopsis cell suspension. *Plant Mol. Biol.*, **53**, 423–442.
16. Gauthier,N., Larsen,M.E., Wernersson,R., de Lichtenberg,U., Jensen,L.J., Brunak,S. and Jensen,T.S. (2008) Cyclebase.org-a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.*, **36**, D854–D859.
17. Caretta-Cartozo,C., de los Rios,P., Piazza,F. and Lió,P. (2007) Bottleneck genes and community structure in the cell cycle network of S. pombe. *PLoS Comput. Biol.*, **3**, 968–976.
18. Moura,D., Castilhos,B., Immich,B., Cañedo,A.D., Henriques,J.A.P., Lenz,G. and Saffi,J. (2010) Kin3 protein, a NIMA-related kinase of Saccharomyces cerevisiae, is involved in DNA adduct damage response. *Cell cycle*, **9**, 2220–2229.
19. Liu,D., Umbach,D.M., Peddada,S.D., Li,L., Crockett,P.W. and Weinberg,C.R. (2004) A random-periods model for expression of cell-cycle genes. *Proc. Natl Acad. Sci. USA*, **101**, 7240–7245.
20. Rueda,C., Fernández,M. and Peddada,S.D. (2009) Estimation of parameters subject to order restriction on a circle with application to estimation of phase angles of cell-cycle genes. *J. Am. Stat. Assoc.*, **104**, 338–347.
21. Jothi,R., Przytycka,T. and Aravind,L. (2007) Discovering functional linkages and uncharacterized cellular pathways using

phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**, 173.
22. Bushel,P., Heard,N., Gutman,R., Liu,L., Peddada,S. and Pyne,S. (2009) Dissecting the fission yeast regulatory network reveals phase-specific control elements of its cell cycle. *BMC Syst. Biol.*, **3**, 93.
23. Liu,J., Smith,C., DeRyckere,D., DeAngelis,K., Martin,G. and Berger,J. (2000) Structure and function of Cdc6/Cdc18: implications for origin recognition and check point control. *Mol. Cell*, **6**, 637–648.
24. Rhind,N. and Russell,P. (2001) Roles of the mitotic inhibitors Wee1 and Mik1 in the G2 DNA damage and replication check points. *Mol. Cell Biol.*, **21**, 1499–1508.
25. Wang,B., Feng,L., Hu,Y., Huang,S.H., Reynolds,C.P., Wu,L. and Jong,A.Y. (1999) The essential role of Saccharomyces cerevisiae CDC6 nucleotide-binding site in cell growth, DNA synthesis, and Orc1 association. *J. Biol. Chem.*, **274**, 8291–8298.
26. Bartholomew,D.J. (1961) A test of homogeneity for means under restricted alternatives. *J. R. Stat. Soc. Ser. B*, **23**, 239–281.
27. Robertson,T. and Wegman,E.J. (1978) Likelihood ratio tests for order restrictions for exponential families. *Ann. Stat.*, **6**, 485–505.
28. Iverson,G.J. and Harp,S.A. (1987) A conditional likelihood ratio test for order restrictions in exponential families. *Math. Soc. Sci.*, **14**, 141–159.
29. Menéndez,J.A., Rueda,C. and Salvador,B. (1991) Conditional test for testing a face of the tree order cone. *Commun. Stat. Simul. Comput.*, **20**, 751–762.
30. Mardia,K. and Jupp,P. (2000) *Directional Statistics*. John-Wiley & Sons, Chichester.
31. Robertson,T., Wright,F.T. and Dykstra,R.L. (1988) *Order Restricted Statistical Inference*. Wiley, New York.