Merging microarray cell synchronization experiments through curve alignment

Filip Hermans[1,2] and Elena Tsiporkova[1,*]

[1]Computational Biology Division, Department of Plant Systems Biology, Flanders Institute for Biotechnology, Technologiepark 917, 9052 Ghent, Belgium and [2]Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 914, 9052 Ghent, Belgium

ABSTRACT

Motivation: The validity of periodic cell cycle regulation studies in plants is seriously compromised by the relatively poor quality of cell synchrony that is achieved for plant suspension cultures in comparison to yeast and mammals. The present state-of-the-art plant synchronization techniques cannot offer a complete cell cycle coverage and moreover a considerable loss of cell synchrony may occur toward the end of the sampling. One possible solution is to consider combining multiple datasets, produced by different synchronization techniques and thus covering different phases of the cell cycle, in order to arrive at a better cell cycle coverage.

Results: We propose a method that enables pasting expression profiles from different plant cell synchronization experiments and results in an expression curve that spans more than one cell cycle. The optimal pasting overlap is determined via a dynamic time warping alignment. Consequently, the different expression time series are merged together by aggregating the corresponding expression values lying within the overlap area. We demonstrate that the periodic analysis of the merged expression profiles produces more reliable *p*-values for periodicity. Subsequent Gene Ontology analysis of the results confirms that merging synchronization experiments is a more robust strategy for the selection of potentially periodic genes. Additional validation of the proposed algorithm on yeast data is also presented.

Availability: Results, benchmark sets and scripts are freely available at our website: http://www.psb.ugent.be/cbd/publications.php

Contact: elena.tsiporkova@ugent.be, fiher@psb.ugent.be

1 INTRODUCTION

Microarray profiling of highly synchronized cell cultures has been widely employed in recent years for the identification of genes which are periodically regulated during the cell cycle. Such studies have turned out to be particularly successful for yeast and mammals (Spellman *et al*., 1998; Cho *et al*., 2001; Shedden and Cooper, 2002a,b; Whitfield *et al*., 2002; Rustici *et al*., 2004; Peng *et al*., 2005; Oliva *et al*., 2005), where a relatively high degree of cell synchronization can be achieved that lasts for multiple cell cycles. In contrast, periodic cell cycle regulation in plants has not been that exhaustively explored, mainly due to the difficulties in achieving a satisfactory degree of synchronization in plant cell suspension cultures, and consequently, the inability of completing one full cell cycle.

There are two commonly used cell suspensions in plants: tobacco BY2 and *Arabidopsis thaliana* cells. Breyne *et al*. (1999) performed a successful synchronization of the tobacco BY2 cell culture and a subsequent cDNA–AFLP genome-wide expression analysis led to the identification of 1340 periodically expressed genes. However, the sequencing of the tobacco genome is not completed yet and the efforts required for carrying out such studies limit considerably their use for a wide-scale analysis of cell cycle regulation. Menges and Murray (Menges and Murray, 2002a) developed *Arabidopsis* cell suspensions, potentially suitable for synchronization with two alternative methods, aphidicolin block/release or removing and re-supplying sucrose to the growth media. In a subsequent study, (Menges *et al*., 2003), subjected samples of aphidicolin-synchronized cells and of sucrose-starved cells to transcript profiling with Affymetrix microarrays and identified >1000 genes as cell cycle regulated.

However, the validity of such studies is seriously compromised by the relatively poor quality of cell synchrony that can be achieved at present for plants in comparison to yeast and mammals. Neither the tobacco BY2 cell culture, referred to in the plant community as highly synchronizable, nor the *Arabidopsis* cell suspensions could generate cell synchrony persisting beyond one complete cell cycle. In reality hardly 80–90% of one cycle could be covered and a considerable loss of cell synchrony occurred toward the end of the sampling. Unfortunately, the present state-of-the-art plant synchronization techniques cannot offer a better cell cycle coverage. One possible solution is to consider merging multiple datasets produced by different synchronization techniques in order to arrive at a better cell cycle coverage.

In this contribution, we describe a method that enables pasting expression profiles from different plant cell synchronization experiments and results in an expression curve that spans more than one cell cycle. Initially, several sets of genes with well known and experimentally confirmed cell cycle involvement and/or regulation are identified. These are subsequently used to determine the optimal pasting overlap for the entire dataset. Consequently, the different expression time series are merged together by aggregating the corresponding expression values lying within the overlap area.

The optimal pasting overlap is determined via a dynamic time warping (DTW) alignment. The DTW alignment algorithm was developed originally for speech recognition (Sakoe and Chiba, 1978; Sankoff and Kruskal, 1983), and it aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match (according to a suitable metric) between the two sequences is found. Thus, the DTW is a much more robust distance measure for time series than classical distance metrics as Euclidean or a variation thereof since it allows similar shapes to match even if they are out of phase in the time axis. The DTW

[*]To whom correspondence should be addressed.

alignments in this work are performed with our in-house gene time expression DTW warping tool *GenTχWarper* (Criel and Tsiporkova, 2006), a Java-based program implementing the original symmetric DTW algorithm (Sakoe and Chiba, 1978) and providing some additional features, as for instance the possibility for defining an offset and thus performing partial alignments by sliding the time series against each other along the time axis.

The identification of periodically expressed genes employs a permutation-based method, which utilizes a combination of a p-value for regulation and a p-value for periodicity.

2 METHODS

2.1 DTW alignment algorithm

Let us first summarize the important features of the original symmetric DTW algorithm as proposed by Sakoe and Chiba, 1978. Consider two sequences of feature vectors $A = [a_1, a_2, \ldots, a_n]$ and $B = [b_1, b_2, \ldots, b_m]$. These can be arranged on the sides of a grid, with one on the top and the other on the left hand side (Fig. 1). Both sequences start at the bottom left of the grid. Inside each cell a distance measure can be placed, comparing the corresponding elements of the two sequences. To find the best match or alignment between these two sequences one needs to find a path through the grid $P = p_1, \ldots, p_s, \ldots, p_k$ [where $p_s = (i_s, j_s)$], referred to as the warping function, which minimizes the total distance between A and B (Fig. 1). Thus, the procedure for finding the best alignment between A and B involves finding all possible routes through the grid and for each one compute the overall distance, which is defined as the sum of the distances between the individual elements on the warping path. Consequently, the final DTW distance between A and B is the minimum overall distance over all possible warping paths:

$$\mathrm{dtw}(A,B) = \frac{1}{n+m} \min_P \left(\sum_{s=1}^{k} d(i_s, j_s) \right).$$

It is apparent that for any pair of considerably long sequences the number of possible paths through the grid will be very large. The major optimizations or constraints of the DTW algorithm arise from the following observations on the nature of acceptable paths through the grid:

- *Monotonic condition*: $i_{s-1} \le i_s$ and $j_{s-1} \le j_s$, i.e. the alignment path will not turn back on itself. Both the i and j indexes either stay the same or increase, they never decrease.
- *Continuity condition*: $i_s - i_{s-1} \le 1$ and $j_s - j_{s-1} \le 1$, i.e. the path advances one step at a time. Both i and j can only increase by at most 1 on each step along the path.
- *Boundary condition*: $i_1 = 1, i_k = n$ and $j_1 = 1, j_k = m$, i.e. the path starts at the bottom left and ends at the top right.

The foregoing constraints allow to restrict the moves that can be made from any point in the path and so limit the number of paths that need to be considered. The power of the DTW algorithm resides in the fact that instead of finding all possible routes through the grid which satisfy the above conditions, the DTW algorithm makes use of dynamic programming and works by keeping track of the cost of the best path at each point in the grid:

$$\begin{aligned}
\gamma(1,1) &= d(1,1)\\
\gamma(i,1) &= d(i,1) + \gamma(i-1,1)\\
\gamma(1,j) &= d(1,j) + \gamma(1,j-1)\\
\gamma(i,j) &= d(i,j) + \min(\gamma(i,j-1), \gamma(i-1,j-1), \gamma(i-1,j)).
\end{aligned}$$

Consequently, $\mathrm{dtw}(A, B) = \gamma(n, m)/(n + m)$. During the calculation process of the DTW grid, it is not actually known which path minimizes the overall distance, but this can be traced back when the end point is reached.

Our DTW warping tool *GenTχWarper* (Criel and Tsiporkova, 2006) implements the classic DTW algorithm as described in this section. In addition, we have extended the core algorithm with several new features: data adjustment, metric, warping window, offset and anchor point. The most important one for our present task is the possibility for applying an offset, i.e. for performing partial alignments by sliding the time series against each other along the time axis. To our knowledge such a feature has never been used before in combination with the DTW algorithm.

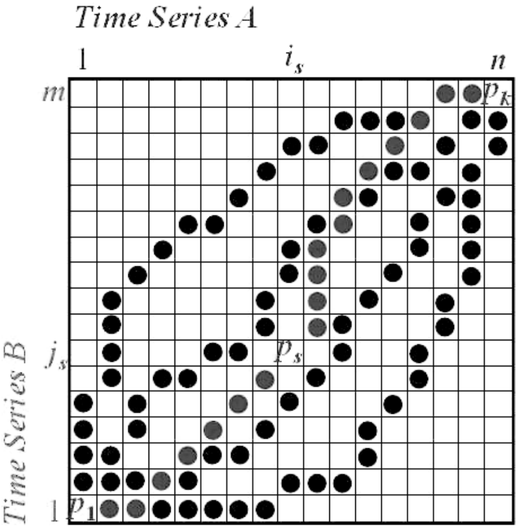


Fig. 1. The DTW grid with different warping paths through it. The optimal warping path is depicted in grey.

In order to facilitate the correct interpretation of our DTW alignments for non-zero offsets, let us give a brief description of the essential details of the offset implementation (Fig. 2), since the latter has not been published elsewhere. It is assumed that a virtual sliding window is positioned on the top of the DTW grid in such a way that the left bottom corners of the DTW grid and the sliding window coincide. The sliding window is a square of a size the length of the longer time series on the DTW grid. Applying a positive offset then corresponds to sliding the window with the offset value at the same time to the left and up from the left bottom corner of the DTW grid. The intersection between the DTW grid and the sliding window determines the new DTW grid. This effectively results in shifting the time series, which is on the left hand side of the DTW grid, to the left with respect to the one on the top of the grid (see the middle graph of Fig. 2). Analogously, a negative offset will mean that the sliding window moves with the offset value right and down, which in its turn causes that the time series on the left hand side of the DTW grid is slided to the right with respect to the one on the top of the grid (see the most right plot in Fig. 2).

2.2 Merging expression profiles from different plant cell synchronization experiments

One possible way to overcome the limitation of incomplete cell cycle coverage imposed by the poor synchronization in plant suspensions is to paste together the expression profiles from synchronization experiments that cover different parts of the cell cycle. We propose to initially determine the optimal pasting overlap between the experiments via a DTW alignment using robust cell cycle genes, and consequently, merge the different expression time series together by aggregating the corresponding expression values lying within the overlap area.

Initially, several sets of genes with well known and experimentally confirmed cell cycle involvement and/or regulation need to be identified. These will be used to determine the optimal pasting overlap for the entire dataset. Then a pairwise DTW alignment of the time series coming from the different synchronization experiments can be performed for the selected set

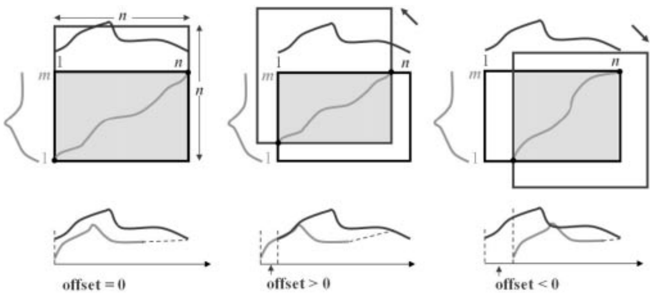


Fig. 2. Illustration of the essential implementation details behind the offset feature. The left most figure corresponds to a classical DTW alignment, i.e. no offset applied. The middle and the right graphs show how the new DTW grid is determined, as the intersection of the original DTW grid and a virtual sliding window (nxn square). The latter is shifted left and up for a positive offset (in the middle) and right and down for a negative offset (most right).

of genes and for a varying offset value. Applying an offset means sliding the time series against each other along the time axis. Due to the fact that cell cycle progression is a periodic process, any pair of cell synchronized expression time series can be aligned against each other in the following two ways:

(1) *With a positive offset*: the expression profiles corresponding to the first synchronization experiment are shifted to the right of time zero of the profiles from the second experiment;

(2) *With a negative offset*: the expression profiles corresponding to the second synchronization experiment are slided to the right of time zero of the profiles from the first experiment.

The range of offset values for which the DTW alignment is performed could be determined from some prior information about the possible time shift between the different synchronizations. The DTW alignment is performed only on the parts of the profiles that overlap.

Finally, for any pair of synchronization experiments, a pair of optimal offset parameters (positive and negative) corresponding to the lowest normalized DTW distance will be determined. Each optimal offset parameter will entail a specific DTW alignment between the two differently synchronized groups of expression profiles. The DTW alignment obtained for the union of all selected cell cycle-associated gene sets is the one that is applied on the entire dataset. Subsequently, the corresponding expression values of the overlapping regions are merged together via a weighted mean aggregation. Each synchronized dataset can be assigned a weight reflecting the quality of synchronization or the degree of cell cycle coverage.

Bear in mind that, the DTW alignments discussed above need to be preceded by some data standardization, as for instance z-transform or $\log_2$-transform. The different expression time series have been generated in different experimental conditions and therefore the comparison of their absolute expression values will not be very meaningful.

2.3 Identification of periodically expressed genes

In a recent study de Lichtenberg *et al*. 2004, have used *Saccharomyces cerevisiae* expression data for benchmarking several computational methods for the identification of periodically expressed genes. In addition to already published methods, they have also proposed a new permutation-based method quantifying separately both the periodicity and the amplitude of variation, and have shown that amplitude-dependent methods perform better than the amplitude independent ones. Taking into account these findings, we consider here a method for the identification of periodically expressed genes, which utilizes a combination of a p-value for regulation and a p-value for periodicity.

The p-value for regulation, referred to as p_{reg}, has been obtained as described by de Lichtenberg *et al*. 2004. Namely, a p-value for regulation for a particular gene is resulting from the comparison of the gene expression variance with a randomly generated variance distribution, constructed by selecting at each time point the log ratio value of a randomly chosen gene.

The p-value for periodicity is obtained in a similar way. As previous microarray analysis has shown (Shedden and Cooper, 2002a), periodic expression patterns can arise from random fluctuations. In order to exclude such cases from the list of genes identified as periodic, a set of artificial expression profiles is generated for each gene by permuting the time points in a random way. Thus, the variances of the artificial expression profiles remain unchanged with respect to the variance of the original gene expression profile. Consequently, some periodicity score is estimated for each observed gene expression profile and then compared with the periodicity scores resulting from the random permutations of the same gene. The p-value for periodicity is calculated as the fraction of artificial profiles with periodicity score equal to or greater than the score of the real expression profile.

The periodicity score is based on a slightly adapted version of the method used by Menges *et al*. 2002b, 2003, described originally by Shedden and Cooper, 2002a. The observed expression profile of each gene i is fit to a periodic component consisting of a sine, a cosine and an amplitude offset:

$$Z_i(t) = a_i S(t) + b_i C(t) + c_i,$$

where $S(t) = \sin(2\pi t/T)$, $C(t) = \cos(2\pi t/T)$ and T is the assumed cell cycle period. The parameters a_i, b_i and c_i can be estimated by means of a linear least squares procedure. Consequently, each gene expression profile will be decomposed into $Y_i(t) = Z_i(t) + R_i(t)$, where $R_i(t)$ represents the component of expression that is either aperiodic or that has a period substantially different from T. Then the ratio $\text{PVE}_i = \text{var}(Z_i(t))/\text{var}(Y_i(t))$, referred to as the proportion of variance explained by the Fourier basis, determines an estimation for periodicity ranging from 0 to 1. The p-values for periodicity derived with the PVE score will be referred to as p_{per}.

As already mentioned above, the latter method for periodicity estimation is a variation of the one used by Menges *et al*. 2002b, 2003. The new element is the introduction of an amplitude (vertical) offset parameter to the periodicity component Z_i. The underlying motivation is that when estimating the periodicity of an expression profile one is only interested in the shape of the curve and not in its absolute position and therefore the regression procedure should be robust against vertical offset. This can be partially achieved by applying some kind of normalization. The only exact vertical translation, however, is the one where the mean of the fitted sine curve is (approximately) the same as the mean of the normalized data. This translation is non-zero for synchronized plant expression data even after performing z-transformation on the original data due to the fact that not an entire (multitude of) period(s) is covered. Hence, the introduction of an explicit amplitude offset parameter makes the sine and cosine parameters independent of vertical translations as long as the regression procedure is linear in its coefficients. In fact, it can be shown that the periodicity score will be invariant for all transformations that are member of the two-parameter (α, β) family $\vec{X}' = \alpha\vec{X} + \beta$.

In the benchmark study of de Lichtenberg *et al*. 2004, a combined p-value was obtained by simply multiplying the p-value for regulation and the p-value for periodicity. This definition entails the negative side effect that the total p-value could become very low due to only one of the individual p-values. Therefore, de Lichtenberg *et al*. (de Lichtenberg *et al*., 2004) had to further introduce two, not really intuitive, penalty terms. We have taken a more straightforward approach to combine the individual p-values for regulation and periodicity, namely through their geometric mean $\sqrt{p_{\text{reg}} \cdot p_{\text{per}}}$, with values always ranging between $\min(p_{\text{reg}}, p_{\text{per}})$ and $\max(p_{\text{reg}}, p_{\text{per}})$. Thus, the combined p-value could be seen as a sort of trade-off between the individual p-values.

Two separate significance conditions need to be verified. For a given significance threshold thr and an individual significance trade-off $\lambda \geq 1$:

$$\sqrt{p_{\text{reg}} \cdot p_{\text{per}}} < \text{thr} \quad \text{and} \quad \max(p_{\text{reg}}, p_{\text{per}}) < \lambda \cdot \text{thr}.$$

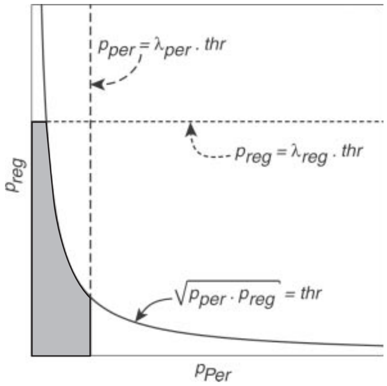

$p_{per} = \lambda_{per} \cdot thr$
$p_{reg} = \lambda_{reg} \cdot thr$
$\sqrt{p_{per} \cdot p_{reg}} = thr$
p_{reg}
p_{Per}

Fig. 3. Graphical illustration of the set of significance conditions.

Consequently, a gene will be qualified as a significantly periodic if the *p*-values associated with it fulfill both conditions. The parameter λ is called an individual trade-off since it determines the degree to which one is prepared to tolerate a violation of the significance threshold by one of the individual *p*-values (either for regulation or for periodicity) as a compensation for a very significant second individual *p*-value. The latter can be refined further by introducing two separate parameters λ_{reg} and λ_{per}, one for each of the individual *p*-values (Fig. 3).

3 RESULTS

In this section, we evaluate the added value of combining time series microarray experiments, originating from different synchronization methods, for a more accurate identification of periodically regulated genes. This issue is particularly important for cell cycle studies in plants, considering that plant sychronization data are characterized with a relatively low cell synchrony which, moreover, persists for less than one complete cell cycle.

The microarray pasting algorithm is applied on two different *Arabidopsis* synchronization expression datasets, and consequently, a set of periodically regulated genes is identified from the combined dataset. An obvious choice of benchmarking data suitable for testing our pasting algorithm is the *Arabidopsis* gene expression data of Menges *et al*. (2003), since it is at present the only available genome-wide synchronization data in plants. The *Arabidopsis* cell suspensions were synchronized with two alternative methods, aphidicolin block/release and sucrose starvation. The latter method produced a partial synchrony from G_0/G_1 until S phase, with some synchrony persisting until mitosis, while with the former a further enhancement of synchrony in the S–G_2–M phases was achieved. Subsequently, 7 samples of sucrose-starved cells and 10 of aphidicolin-synchronized cells, both taken at 2 h intervals starting at time 0, were subjected to transcript profiling with Affymetrix microarrays.

The raw data were downloaded [http://www.arabidopsis.org (submission numbers ME00365 and ME00366)] and RMA preprocessed in Bioconductor (http://www.bioconductor.org). Consequently, *p*-values for regulation and *p*-values for periodicity were calculated as described in Section 2.3.

3.1 Determining the optimal overlap

In order to determine the optimal pasting overlap between the aphidicolin-synchronized expression profiles and the sucrose-starved ones, five different sets of genes with well known cell cycle association were composed. These are B-type cyclins, A-type cyclins, all cyclins, all histones, and all cyclins and histones together. The reasoning behind such choice is the fact that the histones are known to have very consistent behavior during the cell cycle. They are often used as markers for the S phase and it is naturally to expected that their time expression profiles are very conserved. The B-type cyclins, on the other hand, are known to have a peak of transcription during the G_2 to M phase transition and it is believed that they are probably responsible for the mitotic events in plants (De Veylder *et al*., 2003). Although the expression of the B-type cyclins may be highly fluctuating in the cell cycle, they all are expected to have a distinctive peak in early mitosis.

Table 1. *Arabidopsis* (aphidicolin versus sucrose synchronization): normalized DTW alignment scores for different sets of genes and different offset values

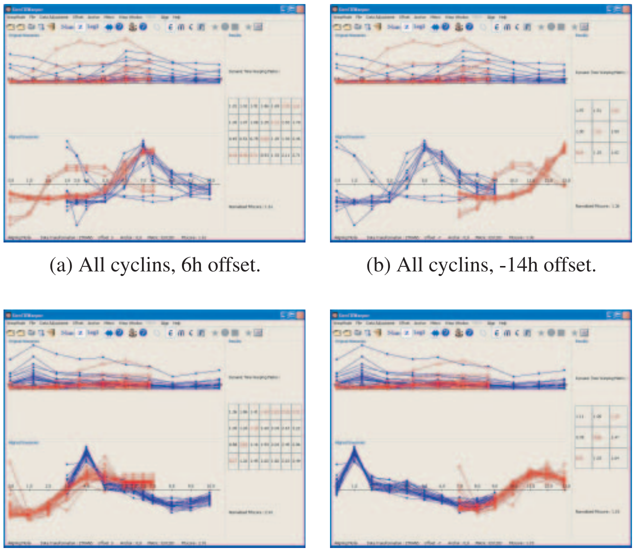
Offset	B-cyclines	A-cyclines	All cyclines	Histones	All
−16 h	**0.68**	0.78	3.26	1.36	2.18
−14 h	1	**0.68**	**1.36**	**1.15**	**1.79**
−12 h	1.49	1.01	1.9	1.37	2.43
−10 h	1.82	1.48	2.39	2.32	3.63
−0 h	1.12	1.69	2.21	4.14	5.25
−2 h	1.05	1.46	2.10	4.03	4.91
−4 h	0.99	1.14	1.78	3.54	4.28
−6 h	**0.82**	**0.98**	**1.61**	**2.91**	**3.60**
−8 h	1.12	1.10	1.81	2.94	3.86

All alignments were performed on *z*-transformed expression profiles. Boldface table entries denote the lowest DTW scores obtained for each dataset.

The range of the offset values for which the DTW alignment was performed was determined from prior information about the possible time shift between the two synchronization methods. For instance, the aphidicolin block/release method causes a synchronous resumption of S phase, while the sucrose starvation results in synchronous transit of G_1 and entry into the first S phase. Thus, there must be a time shift of ~3 to 7 h between the start of the two experiments. Analogously, the sucrose-starved cells were sampled for 12 h until S/G_2 phase, while the aphidicolin-synchronized ones were sampled for a longer period of 19 h until M/G_1 phase. The cell cycle period was estimated at ~22 h by flow cytometry of synchronized cells. Therefore, one can derive that an ~10–16 h time shift exists between the two experiments at the end of the sampling. Consequently, for the five gene sets, defined above, the aphidicolin-synchronized expression profiles were aligned against the sucrose-starved ones with the DTW algorithm for different offset values ranging from −16 to 8 h. The normalized DTW alignment scores of these alignments are reported in Table 1.

Figure 4 presents the DTW alignments, obtained with *GenTχWarper* (Criel and Tsiporkova, 2006), of aphidicolin versus sucrose expression profiles for cyclins and histones, respectively. These correspond to the optimal (corresponding to the lowest DTW score) positive and negative offsets.

Note that in both Figure 4b and d a second peak of expression can clearly be detected. In addition, the DTW scores in Table 1 also indicate that better alignment is achieved when the sucrose expression profiles are pasted after the aphidicolin ones, i.e. when applying a negative offset 14 h. This is not surprising since the aphidicolin cell synchronization is superior to the sucrose starvation one. Thus,



(a) All cyclins, 6h offset.

(b) All cyclins, -14h offset.

(c) Histones, 6h offset.

(d) Histones, -14h offset.

Fig. 4. The original aphidicolin-synchronized (blue) versus sucrose-starved (red) expression profiles are presented in the top panel. Their counterparts *z*-transformed and aligned with the DTW algorithm are visualized in the bottom panel.

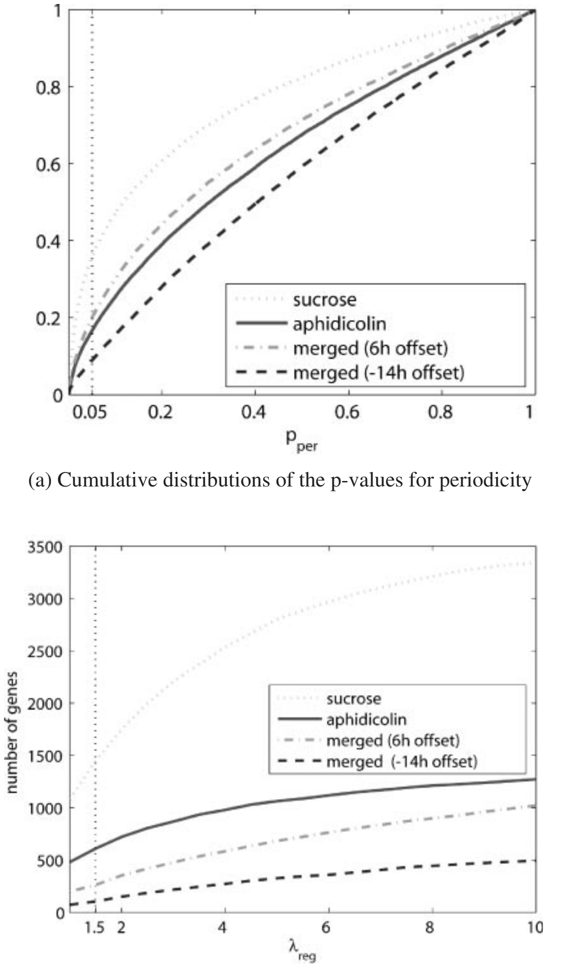
the sucrose-starved profiles are best pasted at the end of the aphidicolin ones, where the sucrose level of synchrony better matches the aphidicolin one, since the latter degrades considerably toward the end of the sampling. The achieved cell cycle coverage for offset -14 h is 26.5 h, i.e. $\sim$1.2 of a cycle.

The best DTW alignment of the sucrose profiles positioned before the aphidicolin ones (Fig. 4a and c), i.e. with a positive offset 6 h, is attained with a much larger overlap than for a negative offset. Despite this larger overlap, this still results in a 22 h cell cycle coverage (1.0 cycle).

3.2 Merging expression profiles from different experiments

A pair of optimal alignment offsets (positive and negative) for the aphidicolin and the sucrose expression profiles was determined from the DTW scores in Table 1. Thus, the two sets of expression profiles could be merged together in two different ways: (1) with a positive offset 6 h or (2) with a negative offset 14 h. Each of these offsets entails a specific DTW alignment between the two groups of expression profiles. The DTW alignments obtained for the union of all selected cell cycle-associated gene sets, i.e. the last dataset in Table 1, are the ones that were applied on the entire dataset. These can be consulted on the web site http://www.psb.ugent.be/cbd/publications.php. The optimal alignment corresponding to an offset -14 h does not involve time warping, while for an offset 6 h multiple warping of the time axis is required. Subsequently, the corresponding expression values of the overlapping regions were merged together via a weighted mean aggregation. Each synchronized dataset was assigned a weight reflecting the extent of cell cycle coverage achieved, i.e. 0.61 (19/31) for aphidicolin and 0.39 (12/31) for sucrose.

In summary, two new merged expression datasets were constructed, one corresponding to a positive offset of 6 h between the experiments and another to a negative offset of 14 h. Subsequently, the combined profiles were assigned *p*-values for regulation equal to the geometric mean of the *p*-values for regulation associated with each of the experiments. The *p*-values for periodicity were estimated as described in Section 2.3.



(a) Cumulative distributions of the p-values for periodicity

(b) The number of genes selected as periodic as a function of λ_{reg} for a significance threshold 0.05 and $\lambda_{per} = 1$.

Fig. 5. Comparison of the results from the periodicity analysis performed on the following four different expression datasets: aphidicolin synchronization, sucrose starvation, aphidicolin and sucrose merged (6 h offset), and aphidicolin and sucrose merged (-14 h offset).

3.3 Identification of periodic genes

Hereafter, we compare the results from the periodicity analysis performed on the four different expression datasets: aphidicolin synchronization, sucrose starvation, aphidicolin and sucrose merged with an offset 6 h, and aphidicolin and sucrose merged with an offset -14 h.

The cumulative distributions of the *p*-values for periodicity for each of the four datasets are given in Figure 5a. The latter clearly illustrates that the higher the extent of cell cycle coverage is, the lower the number of genes with *p*-values for periodicity below a given significance threshold. Thus, for the sucrose-starved profiles, which are covering a little bit more than half of the *Arabidopsis* cell cycle, >35% of all genes have *p*-value for periodicity <0.05. However, for the aphidicolin profiles only ∼18% of the genes have a *p*-value for periodicity below this threshold and in the combined with an offset −14 h dataset hardly 9% such genes are found. The former profiles cover ∼85% of a cycle and the latter >120% of a cycle.

The above phenomenon is due to the fact that an expression profile with a partial cell cycle coverage could end up matching very closely a particular part of a periodic curve, while the same profile over a complete cell cycle is absolutely non-periodic. Actually the observation that in comparison to the original (single experiment) expression profiles, the merged (with an offset −14 h) profiles generate a considerably lower number of genes with *p*-values for periodicity below a given significance threshold (∼50% less than from the aphidicolin profiles) is a clear confirmation that the pasting algorithm has produced meaningful expression profiles. These seem to be a very good approximation of gene expression profiles, as if these were produced by a single synchronization experiment and spanning more than one cycle.

Another way of evaluating the periodicity analysis performance of the four different sets of expression profiles is to compare the number of periodic genes selected from each set as a function of the *p*-value for regulation, according to the two significance conditions as defined in Section 2.3. Figure 5b presents the results of such a comparison for a significance threshold 0.05, $\lambda_{\text{per}} = 1$ and λ_{reg} taking values between 1 and 10. Once again the merged with an offset −14 h profiles exhibit the lowest number of periodic genes. Note that in both plots in Figure 5, the results produced with the merged with an offset 6 h profiles are very comparable with the ones obtained with the aphidicolin profiles. This is probably due to the fact that for this offset value a rather minor extension of the cell cycle coverage, originally obtained with aphidicolin synchronization, is achieved.

Recall that for the merged expression profiles the *p*-values for regulation were obtained as the geometric mean of the individual *p*-values for regulation estimated separately for the original aphidicolin and sucrose profiles. Using the geometric mean guarantees that any gene which is significantly regulated at least in one of the two datasets will also be considered as significantly regulated in the combined dataset. However, this implies that also genes with $p_{\text{reg}} = 1$, i.e. absolutely not significantly expressed and hence having a rather flat expression profile for one of the datasets and p_{reg} close to zero for the other dataset, may also be considered. In order to avoid this, the curves for the merged datasets in Figure 5b were generated by imposing the additional constraint that each individual *p*-value for regulation cannot exceed 0.5. According to the definition of the *p*-value for regulation (see Section 2.3), a *p*-value of 0.5 for a given expression profile means that maximum half of the artificial profiles will have a variance greater than the variance of the real profile. Thus, for a given λ_{reg} a gene will be considered expressed with a significance level $\lambda_{\text{reg}} \cdot 0.05$ in the merged datasets if (1) the gene is significantly expressed in at least one of the experiments; (2) there is at least 50% chance that it is expressed in the other experiment; (3) its merged *p*-value for regulation does not exceed $\lambda_{\text{reg}} \cdot 0.05$.

3.4 Gene Ontology analysis of the results

Figure 6 presents the overlap between the periodic gene sets identified for each of the following four different expression datasets: aphidicolin synchronization, sucrose starvation, aphidicolin and sucrose merged (offset 6 h), and aphidicolin and sucrose merged (offset −14 h). These were obtained for a significance threshold 0.05, $\lambda_{\text{per}} = 1$ and $\lambda_{\text{reg}} = 1.5$. The latter value implies that up to 50% violation of the significance threshold is allowed for the *p*-values for regulation as a trade-off for very significant *p*-values for periodicity.

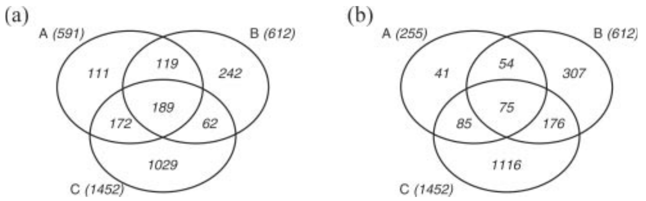


Fig. 6. Set A represents the number of genes identified as periodic from the merged (aphidicolin and sucrose) profiles: (**a**) 6 h offset (**b**) −14 h offset. Set B represents the periodic genes found in the aphidicolin-synchronized profiles and set C in the sucrose-starved ones. All the results are obtained for a significance threshold 0.05, $\lambda_{\text{per}} = 1$ and $\lambda_{\text{reg}} = 1.5$.

Each subset of genes belonging to the Venn diagrams in Figure 6 was subjected to analysis with the BiNGO tool (Maere *et al*., 2005), in order to determine which Gene Ontology (GO) categories are statistically overrepresented in each list. The results for a cutoff *p*-value of 0.05 and using Benjamini and Hochberg (False Discovery Rate) multiple testing correction are presented on our web site http://www.psb.ugent.be/cbd/publications.php. For each gene set a table is generated consisting of four columns: (1) the GO category identification (GO-id); (2) the multiple testing corrected *p*-value (*p*-value); (3) the number of selected genes versus the total GO number (selected/total); and (4) a detailed description of the selected GO categories (description).

The GO categories selected for the intersection of the three different gene sets (aphidicolin, sucrose and merged) show considerable overrepresentation of cell cycle regulation and mitotic control related genes. The same observation can be made for the GO lists for each of the pairwise intersections (aphidicolin/merged, sucrose/merged and aphidicolin/sucrose). However, the triple intersection GO list clearly outperforms the pairwise intersection lists both quantitatively (in terms of gene numbers selected, see column 3 'selected/total') and qualitatively (in terms of *p*-values, see column 2 '*p*-value').

The GO lists of each of the individual gene sets (aphidicolin, sucrose and merged) contain various stress response gene categories as for instance, 'response to abiotic stimulus', 'response to osmotic stress', 'response to wounding', 'response to extracellular stimulus', 'response to salt stress', 'response to oxidative stress', 'response to water deprivation', 'toxin catabolism and metabolism', 'SOS response', etc. Interestingly enough, most of these stress response genes ended up in those gene sets of the Venn diagrams which are specific for each dataset. In fact, in the class of genes from the merged set that are not shared with the individual aphidicolin or sucrose sets only stress-related genes are overrepresented. The aphidicolin-specific gene lists still contain some cell cycle-related GO categories, but they correspond to a very small number of genes (in the range of tens) versus the high total number of genes (in the range of a few hundreds) in these gene lists.

In summary, the GO analysis indicates that there is a higher certainty that a gene is qualified as periodic if it was identified as such in at least two datasets, i.e. it belongs to one of the gene intersection lists (aphidicolin/merged, sucrose/merged, aphidicolin/sucrose and aphidicolin/sucrose/merged). On the other hand, genes specific for a particular dataset are most probably associated with some stress response phenomena.

3.5 Additional validation in yeast

In order to have a fair estimate of the performance of the pasting algorithm, we need to validate the algorithm on expression data coming from another organism with well established synchronization methods that are able to maintain the synchrony for multiple cell cycles. Subsequently, by downsizing the synchronized expression data from multiple cycle coverage to ~85% of a cycle, the cell cycle coverage of plant synchronization data can be mimicked. One of the obvious choices is *Schizosaccharomyces pombe* (fission yeast) as there are synchronization methods reported that manage to synchronize the cells for up to three cell cycles. We have selected the elutriation A and cdc25 experiments from Oliva *et al*. (2005) which are sampled approximately every 10 min for 406 min, covering 2.6 cell cycles. On our web site http://www.psb.ugent.be/cbd/publications.php, we present the results from the periodic analysis performed on (1) the original 2.6 cell cycle coverage expression profiles; (2) expression profiles that have been downsized to 85% cell cycle coverage; and (3) expression profiles resulting from merging data coming from two different synchronization experiments.

The accuracy of periodicity identification on the merged expression profiles has been validated on two different benchmark sets. Following the procedures described in Section 2.3, the respective set of periodic genes for each of the full (2.6 cell cycle coverage) elutriation A and cdc25 experiments has initially been identified. Consequently, two different benchmark sets of ‘truly’ periodic genes have been constructed as follows: (1) the intersection of the elutriation A and cdc25 periodic gene sets; and (2) the union of the elutriation A and cdc25 periodic gene sets.

According to the results presented on the web site, the periodic gene list coming from the merged profiles exhibits the lowest false positive rate for both benchmark sets (2.2% for intersection and 1% for union) in comparison to the lists associated with the elutriation A and cdc25 experiments. Moreover, the merged list also outperforms the elutriation A and cdc25 ones in terms of trade-off between true positives and false positives. This is a very important feature in the context of periodicity studies, which usually target identification of potentially novel cell cycle genes. Ultimately, the hypotheses generated in such studies need to be validated experimentally. The latter process can seriously be hampered by very high false positive rates.

4 CONCLUSION

We have proposed a method that is able of pasting datasets from different synchronization experiments together. As long as synchronization methods in plants will not be able to keep up the synchrony beyond one cell cycle, methods like the one we describe here will probably be the only way to identify periodically regulated genes with some degree of certainty.

Note that the pasting algorithm presented here can easily be extended to the unification of more than two datasets. For instance, one can design and perform a set of multiple synchronization experiments, which most optimally cover all the different phases of the cell cycle. This will ultimately give us a possibility to span several cell cycle periods. Another advantage is that the number of genes that are picked up because of synchronization specific stress responses will diminish.

In conclusion, combining different cell synchronization microarray sets is far from trivial task. However, it may turn out to be useful in a few aspects as follows: (1) provide additional evidence for cell cycle regulation, and consequently, lead to a reduction of the number of false positives; (2) shed light on genes involved in stress response phenomena.

REFERENCES

Breyne,P. *et al*. (1999) Transcriptome analysis during cell division in plants. *Proc. Natl Acad. Sci. USA*, **99**, 14825–14830.

Cho,R.J. *et al*. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet*., **27**, 48–54.

Criel,J. and Tsiporkova,E. (2006) Gene time expression Warper: a tool for alignment, template matching and visualization of gene expression time series. *Bioinformatics*, **22**, 251–252.

de Lichtenberg,U. *et al*. (2004) Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, **21**, 1164–1171.

De Veylder,L. *et al*. (2003) Plant cell cycle transitions. *Curr. Opin. Plant Biol*., **6**, 536–543.

Menges,M. and Murray,J.A.H. (2002a) Synchronous *Arabidopsis* suspension cultures for analysis of cell-cycle gene activity. *Plant J*., **30**, 203–212.

Menges,M. *et al*. (2002b) Cell cycle-regulated gene expression in *Arabidopsis*. *J. Biol. Chem*., **277**, 41987–42002.

Menges,M. *et al*. (2003) Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Mol. Biol*., **53**, 423–442.

Maere,S. *et al*. (2005) BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Oliva,A. *et al*. (2005) The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PloS Biol*., **3**, 1239–1260.

Peng,X. *et al*. (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet*., **36**, 809–817.

Sakoe,H. and Chiba,S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Processing*, **ASSP-26**, 43–49.

Sankoff,D. and Kruskal,J. (1983) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.

Shedden,K. and Cooper,S. (2002a) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *PNAS*, **99**, 4379–4384.

Shedden,K. and Cooper,S. (2002b) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarray and multiple synchronization methods. *Nucleic Acids Res*., **30**, 2920–2929.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Whitfield,M.L. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.