BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 228–233
doi:10.1093/bioinformatics/btt671

Systems biology

Advance Access publication November 18, 2013

Combining optimization and machine learning techniques for genome-wide prediction of human cell cycle-regulated genes

Marianna De Santis[1,†], Francesco Rinaldi[1,†], Emmanuela Falcone[2], Stefano Lucidi[1], Giulia Piaggio[2], Aymone Gurtner[2,*] and Lorenzo Farina[1,*]

[1]Dipartimento di Ingegneria Informatica, Automatica e Gestionale, Sapienza Università di Roma, Rome, Italy and [2]Istituto Nazionale dei Tumori Regina Elena, Dipartimento di Oncologia Sperimentale, Laboratorio di Oncogenesi Molecolare, Rome, Italy

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The identification of cell cycle-regulated genes through the cyclicity of messenger RNAs in genome-wide studies is a difficult task due to the presence of internal and external noise in microarray data. Moreover, the analysis is also complicated by the loss of synchrony occurring in cell cycle experiments, which often results in additional background noise.

Results: To overcome these problems, here we propose the LEON (LEarning and OptimizatioN) algorithm, able to characterize the 'cyclicity degree' of a gene expression time profile using a two-step cascade procedure. The first step identifies a potentially cyclic behavior by means of a Support Vector Machine trained with a reliable set of positive and negative examples. The second step selects those genes having peak timing consistency along two cell cycles by means of a non-linear optimization technique using radial basis functions. To prove the effectiveness of our combined approach, we use recently published human fibroblasts cell cycle data and, performing *in vivo* experiments, we demonstrate that our computational strategy is able not only to confirm well-known cell cycle-regulated genes, but also to predict not yet identified ones.

Availability and implementation: All scripts for implementation can be obtained on request.

Contact: lorenzo.farina@uniroma1.it or gurtner@ifo.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 11, 2013; revised on November 13, 2013; accepted on November 14, 2013

1 INTRODUCTION

Cell cycle progression is achieved by a highly regulated temporal program of gene expression where transcription is coordinated in a series of consecutive interdependent waves. Such 'just-in-time' strategy is suggested by the experimental observation that genes are expressed at peak levels at the time they are needed (Breeden *et al.*, 2003). To study genes regulated in a cell cycle-dependent manner, researchers have deeply used messenger RNA expression profiling microarrays of synchronously growing cells progressing through the cell cycle. Nevertheless, the genome-wide identification of cycling genes is a difficult task for a number of reasons, including cell synchronization loss and intrinsic microarray noise. The most critical experimental issue is the degree of synchronization as well as its loss during a time course sampling. This is the reason why various computational methods have been proposed to correct such experimental artifact. The key problem is that the loss of synchronization results in a 'flattening' of the time profile due to the presence of cells in different stages of the cell cycle. To overcome this problem, Bar-Joseph *et al.* (2008) have recently developed a combined experimental and computational approach to recover 'true' cell cycle expression profiles. A number of different techniques have been proposed in the literature for the identification of cycling genes in many organisms, all of them relying on a cyclicity scores based on the degree of regulation (magnitude) and/or on a shape parameter (Fourier analysis). Unfortunately, there is a remarkably poor overlap between the gene sets identified as cyclic in different experiments even for the same organism (Zhao *et al.*, 2001). Therefore, none of them can be considered as the 'best method' and this field is still an open area of research.

Here, we provide a new tool to identify cycling transcripts by integrating different features of gene expression time profiles. In fact, the LEON (LEarning and OptimizatioN) algorithm combines, in two separate steps, information contained in the amplitude of the response (learning step) and information contained in the shape of the response by considering the consistency of peak timing between minima and maxima along two cell cycles (optimization step). In particular, this last feature is suitable to overcome the loss of synchronization problem, as peak timing of upregulation or downregulation is not significantly affected by such artifact. In other words, the shape of the expression profile of the population of cells changes as a consequence of different cell progress rate, but the up or down expression peak positions in time remain approximately the same. This property is illustrated in Figure 1 through the example used by Bar-Joseph *et al.* (2008) where raw and deconvolved data display the same peak timing. The authors hypothesize that cell progress rates are distributed as a Gaussian with a mean of 1 (average time) and that observed expression values result from a convolution with such Gaussian kernel. Because the Gaussian kernel function is symmetric around the mean, if a certain group of cells has a cell progress rate, say $1 - d$, there is also a 'symmetric' group of cells having a progress rate $1 + d$. The total effect of these contributions leaves the

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

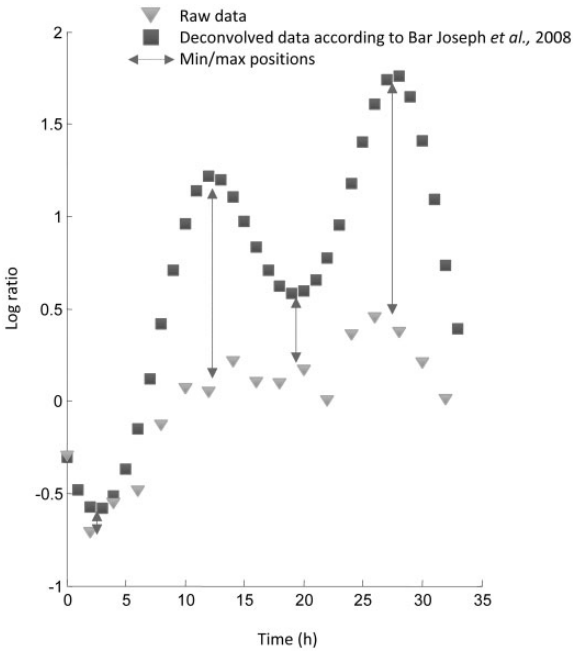

Fig. 1. Experimental gene expression time profile (triangles) and deconvolved data (squares) using BIRC5 (data taken from Bar-Joseph *et al.*, 2008). It is worth noting that peaks remain located approximately at the same position in time both in raw and deconvolved data

maxima points (and the minima points) in the same time positions.

To prove the effectiveness of our approach, we used recently published human fibroblasts cell cycle data (Bar-Joseph *et al.*, 2008) and performed *in vivo* experiments to validate our computational predictions.

2 METHODS

The LEON algorithm identifies cycling genes by a two-step procedure. The first step computes a suitably trained Support Vector Machine (SVM) classifier to decide whether a given profile is cycling or not and, accordingly, a binary cyclicity score of c = 1or c = 0 is assigned. The second step consists in an optimization-based approach, which identifies consistency of peak timing along two cell cycles and accordingly assigns to each gene a cyclicity score p between 0 and 1. The SVM prediction and the optimization procedure are then combined into a single overall score p_{comb} characterizing the degree of cyclicity for each transcript.

2.1 First step: SVM classification

The SVMs, first described by Vapnik in the 90 s (Vapnik, 1998), have been successfully applied to a large number of pattern recognition problems, including classification of microarray data (Brown *et al.*, 2000). Each expression profile vector with n time samples can be seen as a point in an n-dimensional space. To separate in this space cycling genes from non-cycling genes, the simplest approach is to consider a linear decision boundary region (i.e. a hyperplane) that separates the two given classes (positive and negative examples) in the original n-dimensional space. In our case such simple rule is not satisfactory, as the 'best' hyperplane separating the two classes still misclassifies a large number of genes. Therefore, we adopted the common approach of mapping the dataset into a higher dimensional space (the so-called feature space).

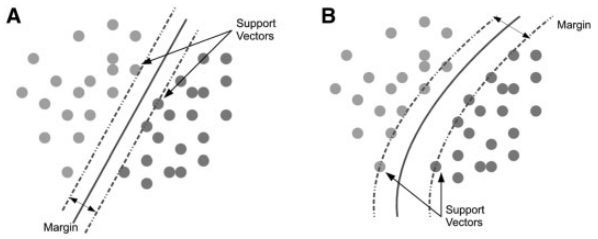

Fig. 2. Examples of SVM decision boundary regions. (**A**) linear and (**B**) non-linear

This point is illustrated in Figure 2. Then, a reliable subset of cycling and non-cycling genes was used as positive and negative examples for the training set. On this basis, the first step of LEON algorithm, classifies all genes as 'cyclic' or 'not cyclic' (see Fig. 3) using an SVM (see Section 3).

2.2 Second step: non-linear optimization-based gene expression profile analysis

Cyclicity of a time series can be characterized in many ways. A common choice is the use of the first Fourier coefficient (Spellman *et al.*, 1998), but often the shape of a time series is different from a pure sinusoid. Additionally, owing to the already mentioned synchronization loss, the shape of the curve is also different between two subsequent cycles of the same transcript.

Among the many possible alternatives, we selected a feature that is independent from the overall shape and robust with respect to the loss of synchronization problem (see Fig. 1), as previously described. Such feature is the distance in time d_{min} between two subsequent minima and the distance d_{max} between two subsequent maxima (illustrated in Fig. 4). Therefore, cell cycle time series having d_{min} and d_{max} near to the duplication time Δ, namely, the time needed by a cell to complete a cell division cycle obtained by flow cytometry [Fluorescence-Activated Cell Sorting (FACS)] analysis, are considered good candidates for being associated to cycling genes.

Unfortunately, microarray time series are noisy and provide expression values only at sampling times. To robustly evaluate peak times, data were preprocessed to reduce noise and then values outside the range of sampling times were extrapolated. By doing so, we could obtain a precise estimation of times at which maxima or minima are attained.

Therefore, we needed an appropriate mathematical tool to get a smooth continuous representation of the experimental data that enables us to estimate unobserved time points. Radial basis functions (RBFs) are tools to approximate functions by linear combinations of terms based on a single univariate function (the RBF).

The use of RBFs is motivated by the fact that they have excellent approximation properties (as shown, e.g. in Girosi and Poggio, 1990) and are widely used also in the bioinformatics field (Chen *et al.*, 2011; Chiang and Ho, 2008; Takasaki *et al.*, 2006). They are usually applied to approximate data that are only known at a limited number of points—as for the case of microarray experiments—so that the approximating function can be evaluated often and efficiently. In fact, once the RBFs are generated, one can resample the curve to estimate expression values at any time points.

The second step of LEON algorithm consists in the approximation and resampling of (noisy) expression time profiles using a linear combination of RBF. When estimating RBFs from expression data, we did not fit each time course individually. Owing to noise and missing values, such an approach could lead to overfitting. Instead, we constrained the RBF coefficients to have good smoothness (or regularity) properties (see Section 3.2) based on the same training set (golden standard genes) used for SVM. Then, a cyclicity score p can be computed for each gene using a non-linear optimization-based method to find optimal parameter values of the RBFs. That is, a set of optimal RBF parameters (the same for all

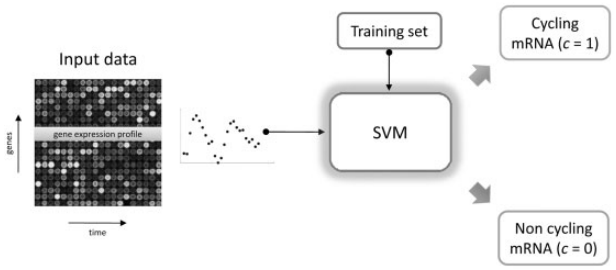

Fig. 3. Gene expression profile classification procedure using an SVM

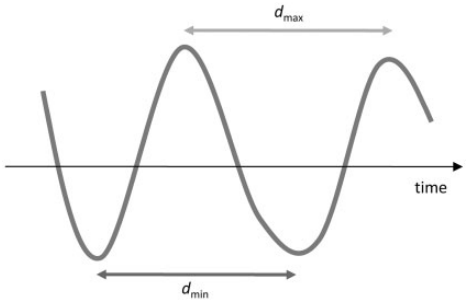

Fig. 4. Time series parameters used to characterize cyclicity

genes) is calculated by means of an optimization-based method. The rationale behind this choice is to determine the set of parameters that best represent the biological process under analysis. Then, for a given gene, we calculated the RBF approximation and executed a cyclicity analysis to get a score p. Precisely, once the smoothed expression profile is computed, the cyclicity score p is defined as follows:

$$p = \frac{m_{max} + m_{min}}{2}$$

$$m_{max} = 1 - \frac{|d_{max} - \Delta|}{\Delta}, \quad m_{min} = 1 - \frac{|d_{min} - \Delta|}{\Delta}$$

where d_{max}, d_{min} are, respectively, the distances in time between the two maximum peaks and the two minimum peaks of the approximation curve and Δ is the duplication period. A score value close to 1 indicates a high consistency of peak timing and, therefore, a good indication of the presence of a cyclic behavior. To find optimal parameters values, we evaluated the performance of the smoothing algorithm by generating a Receiver-Operating Characteristic (ROC) curve according to the p score (Fig. 5). Then, a global optimization algorithm (namely, the PRICE algorithm see e.g. Brachetti *et al.*, 1997; Liuzzi *et al.*, 2003) found the parameters giving the best ROC curve (larger area under curve) obtaining $A.U.C. = 0.86$. The final RBF coefficients were $\tau_1 = 4 \times 10^{-4}$, $\tau_2 = 1 \times 10^{-4}$, $\sigma = 2$, and the number of RBFs used was 8. The overall algorithm scheme is illustrated in Figure 6.

2.3 Combined cyclicity score

In sum, the proposed methodology is the following: first, we decided whether a given gene expression is sufficiently 'fluctuating' (cycling genes usually have such a kind of expression). Then, we identify among those fluctuating genes, the ones that are cycling by means of a peak consistency analysis. In particular, the SVM is used to extract the information contained in the amplitude of fluctuations and to get rid of bad shaped expression profiles (e.g. flat profiles) that usually identify not cycling genes. Then, the p score, which is related to the consistency of peak timing along two cell cycles, identifies those genes that are cycling.

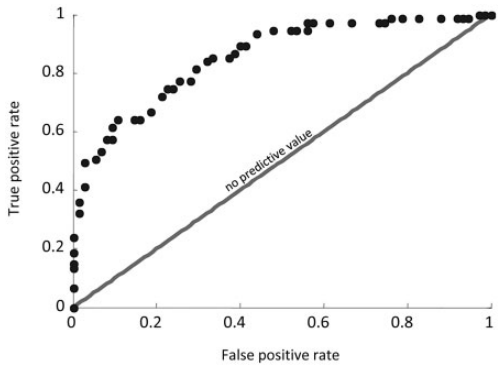

Fig. 5. ROC curve. (A.U.C. = 0.86)

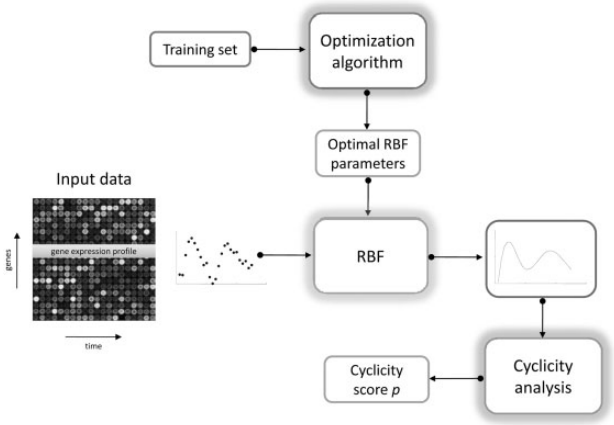

Fig. 6. Optimization-based gene expression cyclicity analysis via RBFs

Once each gene in our test set was classified (c score) by the learning machine (first step) and the p score for that gene was calculated (second step), we built a combined score as follows:

$$p_{comb} = \frac{c + p}{2}$$

where c is the SVM output related to the gene under analysis ($c = 1$ cyclic, $c = 0$ not cyclic). Then we ranked genes according to their p_{comb} value. Ideally, the first gene in the list should be the one classified as cycling ($c = 1$) and with a p_{comb} score equal to 1, whereas the last one should be the one classified as non-cycling ($c = 0$) and with a p_{comb} score equal to 0.

As a preliminary step, we generated synthetic data to evaluate the differentiating ability of the proposed method. To produce realistic synthetic data, we followed the approach of Zhao *et al.* (2001). The authors propose an algorithm to generate synthetic time series by assuming that cell cycle-regulated genes are transcribed at one invariant time and that synchrony deteriorates with time, leading to the attenuation of simple pulses into smooth peaks that dampen out with time. Using this algorithm, we generate 1000 synthetic time courses covering two cell cycles, and 1000 randomly fluctuating profiles, obtained by random time shuffling of cyclic data. Then, we generated, using the same algorithm, 50 positive and 50 negative synthetic examples for the learning step. We also added to all synthetic data a multiplicative white Gaussian noise having 10 and 20% of noise standard deviation. We applied the proposed method to such data and computed ratios of genes scored >0.5 and <0.5 for cyclic (Rc) and non-cyclic (Rnc), respectively, and obtained a ratio of Rc/Rnc = 1/0.997 = 1.003 with 10% of noise and a ratio of Rc/Rnc = 1/0.993 = 1.007

with 20% of noise thus, indicating a high differentiating power on synthetic data, slightly favoring positive examples (cyclic genes).

To test our methodology on real experiments, we used microarray data taken from Bar-Joseph *et al.* (2008), which consist of a population of synchronized foreskin fibroblast cells at 2 h intervals after their release from double-thymidine blockarrest (two cell cycles). The list of all genes and the corresponding scores are reported in Supplementary Table S2. We computed the p_{comb} score of the 480 genes indicated as cyclic by Bar-Joseph *et al.* (2008) and the resulting p_{comb} score distribution is reported in Figure 7. The picture shows a clear bimodal distribution composed of a group of high p_{comb} score (dark bars) and a group of low score (light bars). Two main features of the score distribution (Fig. 7) indicate a good performance of the proposed methodology: first, the low values (flat profile genes) are almost uniformly distributed, thus indicating that the value of score p does not add any information, consistently with the 'flat' nature of the time series revealed by the c score. Second, the high values (fluctuating profile genes) are skewed toward the maximal value 1, thus indicating that the score p actually provides additional information about cyclicity. In other words, the combination of the two scores performs better then each single one.

To further verify LEON performance, we also used an independent source to check cyclicity. We considered the Cyclebase database (Gauthier *et al.*, 2009) where human genes are classified as cyclic using cell cycle expression data of *HeLa* cells (Whitfield *et al*, 2002). Supplementary Table S3 reports, for each gene of the low p_{comb} score group, the annotation and the Cyclebase classification. We found 91 genes, of which 18 are classified as cyclic, 44 non-cyclic and 29 not classified. Therefore, not considering the unclassified genes, we found 71% of the genes in this group classified by Cyclebase as non-cyclic, in good agreement with our analysis. Moreover, among the genes for which a Cyclebase ranking is not available, we considered PRKD1 for *in vivo* validation, and we actually found a non-cyclic pattern (Fig. 8).

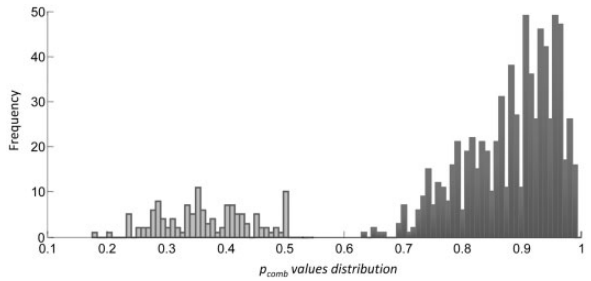

Fig. 7. p_{comb} score distribution for the 480 genes considered as cyclic by Bar-Joseph *et al.* (2008)

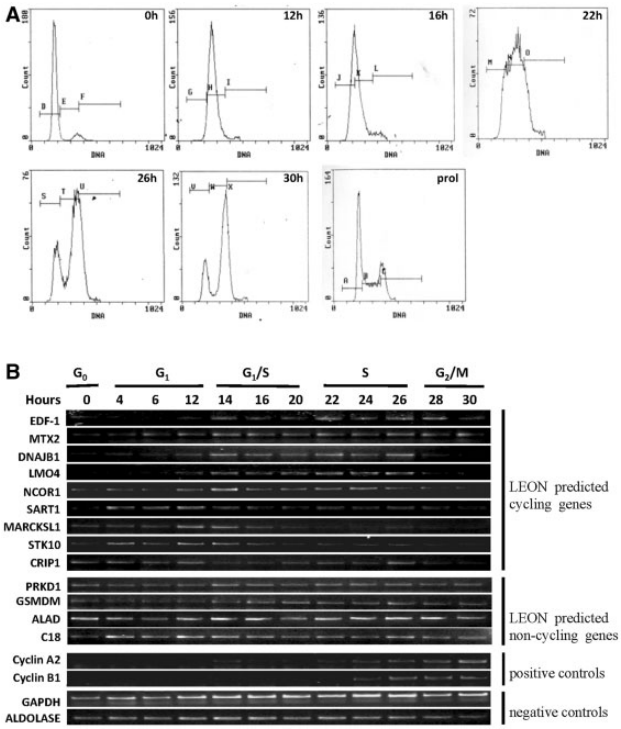

Fig. 8. *In vivo* validation of LEON predicted cell cycle-regulated genes. (**A**) Cell cycle synchronization. The cell cycle distribution of synchronized cells was monitored by FACS. The position of the gates is referred to that in proliferation cells. (**B**) Total RNA was isolated from synchronized cells at the indicated phases of the cell cycle and used for RT-PCR of nine LEON predicted cycling genes and four LEON predicted non-cycling genes. Cyclin A2 and cyclin B1 were used as positive controls. GAPDH and aldolase genes were used as housekeeping constitutive expressed genes (negative control)

3 IMPLEMENTATION

3.1 SVM classification

We selected 50 gene expression profiles as positive examples, compiled from the literature that had been shown to be cell cycle-regulated in cells synchronized experiments (Supplementary Table S1). Known cell cycle-regulated genes were limited to those regulated at the messenger RNA level during a continuous human cell cycle, as determined by traditional experimental methods. We selected also 50 profiles generated by a random time shuffling procedure, considered as negative ones. Eventually, we obtained a sample of 100 genes each one having dimension 17 (time points).

Then, we trained the SVM using the LIBSVM software, a library for SVMs developed by the Machine Learning Group at National Taiwan University (Chang *et al.*, 2008) and selected a radial basis kernel. The parameters C and gamma were determined by a standard k-fold cross-validation procedure (with $k=5$) combined with a grid search. This procedure is done to prevent overfitting problems (Bishop *et al.,* 1996; Hsu *et al.*, 2010). In particular, we first divided the training set into five subsets of equal size. Sequentially one subset was tested using the classifier trained on the remaining four subsets. Thus, each instance of the whole training set was predicted once, so the k-fold cross-validation accuracy is the percentage of data that are correctly classified. Various pairs of C and values (selected from a suitably chosen grid) were tried and the one with the best k-fold cross-validation accuracy was picked. The final values were $C=1$ and $=0.06$. The final k-fold cross-validation accuracy and standard error were, respectively, 93.1%, and 0.02.

3.2 Radial basis approximation

The approximating function $\tilde{y}: R \rightarrow R$used was as follows:

$$\tilde{y}(t) = \sum_{j=1}^{m} a_j \phi(|t - c_j|)$$

where ϕ is a suitably chosen RBF, a_j are the coefficients of the linear expansion and c_j are the centers of the RBFs. We used the inverse multiquadric RBF in our experiments:

$$\phi(r) = \left(r^2 + \sigma^2\right)^{-1/2}$$

where σ is a positive scalar. This choice was mainly motivated by the fact that the multiquadric function showed the best performance in expression profile approximation. To find the approximation, we solved the following regularized problem:

$$\min_{a,c} \sum_{i=1}^{P} (y(t_i) - \tilde{y}(t_i)) + \tau_1 \|a\|^2 + \tau_2 \|c\|^2$$

where $y(t_i)$ is the expression of the gene at the sampling instant t_i. The number of RBFs, the RBF parameter σ and the regularization parameters τ_1 and τ_2 were the same for all the approximations and were obtained by means of a global optimization algorithm. The idea behind this approach is that of determining the values of the parameters that better represent the biological process under analysis. For a given choice of m, σ, τ_1 and τ_2, we computed the expression profile approximations and the corresponding scores p of the genes in the dataset used for the training of the SVM.

3.3 Cell culture, synchronization and FACS analysis

Early passage of human foreskin fibroblasts was grown in Dulbecco's Modified Eagle Medium (DMEM) with 10% Fetal Calf Serum (FCS). For G0/G1 synchronization, cells were arrested with 0.5% FCS (48 h) and then released in 10% FCS. Cells were harvested at specific time points following serum stimulation and were processed. The progression of cells through the cell cycle was monitored by flow cytometric (FACS) analysis of replicate samples of propidium iodide-stained cells (Gurtner *et al.*, 2008).

RNA extraction and reverse transcriptase-polymerase chain reaction (RT-PCR). Total RNA was extracted using the Trizol reagent (Gibco BRL) and following the manufacturer's instructions. The first strand of complementary DNA was synthesized according to manufacture instructions of the M-MLV RT kit (Invitrogen). PCR was performed with HOT-MASTER Taq (Eppendorf) using 2 ml of complementary DNA reaction. The primer sequences of the human genes are as follows:

hCycA2f: 5′-AGCAGCCTGCAAACTGCAAAGTTG
hCycA2r:5′-TGGTGGGTTGAGGAGAGAAACACC
hCycB1f :5′-CCTCTACCTTTGCACTTCCTTCGG
hCycB1r: 5′-GAGTGCTGCTCTTAGCATGCTTCG
hMTX2F:5′-CTGCAGAACCTTGGCCTGAA
hMTX2R :5′-CTGCACTGCAAGAGAAGCTG
hEDF1F:5′-GCACAGAGACGAGGAGAAGA
hEDF1R:5′-ACCTTGCTGGATCACCTTGC
hDNAJF:5′-TCAAGGAGATCGCTGAGGCC
hDNAJR: 5′-GCCCATAGGGAAGCCAGAGA
hLMO4F:5′-GGAAATAGCGGTGCTTGCAG
hLMO4R:5′-GGCAGTAGTGGATTGCTCTG
hNCOR1F:5′-GACCTGACCAATATGCCTCC
hNCOR1R:5′-AAGCTGCAGCAATCCGTTCC
hMARCKSL1F :5′-AGCCAGAGCTCCAAGGCTC
hMARCKSL1R:5′-CTCTTCCTCTGTGGGTGAGG
hSART1F:5′-GTCCAAGAAGCATCGCGGAG
hSART1R:5′-GTAGCCGTCATCGCGCTTCT
hSTK10F:5′-TGCGCCTGTCTACCTTCGAG
hSTK10R:5′-CCTCTTGCTGTGCAGGAAGT
hCRIP1F:5′-CCAAGTGCAACAAGGAGGTG
hCRIP1R:5′-CTTGAAAGTGTGGCTCTCGG
hCDC27F:5′-GGTTTTCCTCGCAGAACGCC
hCDC27R:5′-CCTTTGGCAAGCCATCTGT
RThGSDMDF:5′-GTGGCAGGAGCTTCCACTTC
RThGSDMDR:5′-CCTCAGTCACCACGTACACG
hALADF:5′-GAAGCGGCTGGAAGAGATGC
hALADR:5′-CTCAGCCCGGAATGCTCCGT
hC18F:5′-CATCTGGCAATGCGCCACTC
hC18R:5′-CTGCTGGTGAGCCCAAGTC
hGAPDHF :5′-TCCATGACAACTTTGGCATCGTGG
hGAPDHR:5′-GTTGCTGTTGAAGTCACAGGAGAC
hAldF:5′-CGC AGA AGG GGT CCT GGT GA
hAldR:5′-CAG CTC CTT CTT CTG CTG CG
hPRKD1F:5′-AATGCTGTGGGGGCTGGTAC
hPRKD1R:5′-GTACCAGCCCCCACAGCATT

4 RESULTS

4.1 *In vivo* validation of LEON predicted cell cycle-regulated genes

Using LEON algorithm we selected 50 genes having the highest p_{comb} score and excluded 5 genes that were identified as cycling also by Bar-Joseph *et al.* (2008). Among the remaining 45 new putatively cycling genes, we experimentally validated the cell cycle-dependent expression of 9 of them. To this purpose, primary human fibroblasts prepared from human foreskin were grown to ~50% confluence and synchronized in G_0 by serum deprivation. Cultures were then released from arrest, and cells were collected at different time points covering one complete cell cycle (Fig. 8A).

We used RT-PCR to measure the expression level of the nine genes with higher combined score (see Supplementary Table S2 and Fig. 7) along synchronized cell populations. None of these genes result to be cell cycle-regulated in the study by Bar-Joseph *et al.* (2008), demonstrating the strength of the LEON algorithm in identifying cell cycle-expressed genes. As shown in Figure 8B (RT-PCR) and Supplementary Figure S1 (densitometric analysis), the expression of these nine genes is cell cycle-regulated being maximum on S phase for six and on G1 phase for four of them, respectively.

Interestingly, the expression of two of these genes, NCOR1 and EDF-1, is already known to be cell cycle-regulated further

confirming the predictive power of the LEON algorithm (Altintas *et al*., 2011; Bolognese *et al*., 2006). Next we measured the expression level of four genes with a low combined score. Of note, the expression of these genes is not regulated during the cell cycle further demonstrating that the LEON algorithm specifically identifies the cell cycle gene expression. We measured cyclin A and cyclin B1 as well as GAPDH and aldolase genes as positive and negative controls, respectively. It is well known that cyclin A and B1 expression is cell cycle dependent being high in S phase and in G2, respectively (Pines and Hunter, 1989, 1990). On the contrary, the expression of GAPDH and aldolase genes is constant during the cell cycle (Eisenberg and Levanon, 2003). As expected, cyclin expression is cell cycle dependent, whereas GAPDH and aldolase expression is constant. These results demonstrate that our approach is successful in identifying cell cycle-regulated genes.

Funding: This work has been partially supported by grants from AG and GP AIRC (MFAG 11752 and IG 13234).

Conflict of Interest: none declared.

REFERENCES

Altintas,D.M. *et al*. (2011) Cell cycle regulated expression of NCoR might control cyclic expression of androgen responsive genes in an immortalized prostate cell line. *J. Mol. Cell. Endocrinol*., **30**, 149–162.

Bar-Joseph,Z. *et al*. (2008) Genome-wide trasncriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci. USA*, **22**, 955–960.

Bishop,C.M. (1996) *Neural Networks for Pattern Recognition*. Oxford University Press, New York, USA.

Bolognese,F. *et al*. (2006) Characterization of the human EDF-1 minimal promoter: involvement of NFY and Sp1 in the regulation of basal transcription. *Gene*, **7**, 87–95.

Brachetti,P. (1997) A new version of the Price's algorithm for global optimization. *J. Glob. Optim*., **10**, 165–184.

Breeden,L.L. (2003) Periodic transcription: a cycle within a cycle. *Curr. Biol*., **13**, R31–R38.

Brown,M.P.S. *et al*. (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Chang,K.W. *et al*. (2008) Coordinate descent method for large-scale L_2-loss linear SVM. *J. Mach. Learn. Res*., **9**, 1369–1398.

Chen,S.A. *et al*. (2011) Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties. *Bioinformatics*, **27**, 2062–2067.

Chiang,J.H. and Ho,S.H. (2008) Combination of rough-based feature selection and RBF neural network for classification using gene expression data. *IEEE Trans. Nanobioscience*, **7**, 91–99.

Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet*., **19**, 362–365.

Gauthier,N.P. *et al*. (2009) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res*., **38**, D699–D702.

Girosi,F. and Poggio,T. (1990) Networks and the best approximation property. *Biol. Cybern*., **63**, 169–176.

Gurtner,A. *et al*. (2008) NF-Y dependent epigenetic modifications discriminate between proliferating and postmitotic tissue. *PLoS One*, **3**, e2047.

Hsu,C.W. *et al*. (2010) A practical guide to support vector classification. *Technical report*, Department of Computer Science, National Taiwan University.

Liuzzi,G. *et al*. (2003) Multi-objective optimization techniques for the design of induction motors. *IEEE Trans. Magn*., **39**, 1261–1264.

Pines,J. and Hunter,T. (1989) Isolation of a human cyclin cDNA: evidence for cyclin mRNA and protein regulation in the cell cycle and for interaction with p34cdc2. *Cell*, **58**, 833–846.

Pines,J. and Hunter,T. (1990) Human cyclin A is adenovirus E1A-associated protein p60 and behaves differently from cyclin B. *Nature*, **346**, 760–763.

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.

Takasaki,S. *et al*. (2006) Selecting effective siRNA sequences by using radial basis function network and decision tree learning. *BMC Bioinformatics*, **7** (**Suppl. 5**), S22.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley, New York.

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.

Zhao,L.P. *et al*. (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.