Journal of Theoretical Biology 459 (2018) 130–141

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtb

A feature-based integrated scoring scheme for cell cycle-regulated genes prioritization

Lorenzo Farina[a,b,*], Paola Paci[b,c]

[a] Department of Computer, Control and Management Engineering "A. Ruberti", Sapienza University of Rome, Italy
[b] Institute for Systems Analysis and Computer Science "A. Ruberti", National Research Council, Rome, Italy
[c] SysBio Centre for Systems Biology, Rome, Italy

ARTICLE INFO

Article history:
Received 23 March 2018
Revised 3 August 2018
Accepted 23 September 2018
Available online 24 September 2018

Keywords:
Budding yeast
Gene expression
Time-series
Cell cycle

ABSTRACT

Prioritization of cell cycle-regulated genes from expression time-profiles is still an open problem. The point at issue is the surprisingly poor overlap among ranked lists obtained from different experimental protocols. Instead of developing a general-purpose computational methodology for detecting periodic signals, we focus on the budding yeast mitotic cell cycle. The reason being that the current availability of a total of 12 datasets, produced by 6 independent groups using 4 different synchronization methods, permits a re-analysis and re-consideration of this problem in a more reliable and extensive data domain. Notably, budding yeast is a model organism for studying cancer and testing new drugs. Here we propose a novel multi-feature score (called PERLA, PERiodicity, Regulation and Lag-Autocorrelation) that integrates different features of cell cycle-regulated gene expression time-profiles. We obtained increased performances on a wide range of benchmarks and, most importantly, a substantially increased overlap of the top ranking genes among different datasets, thus proving the effectiveness of the proposed prioritization algorithm. Examples on how to use PERLA to gain new insight into the biology of the cell cycle, are provided in a final dedicated section.

1. Introduction

The issue of prioritizing cell cycle-regulated (CCR) genes using expression time-profile experiments, *i.e.* finding a way to rank them in terms of their role during the cell-cycle, is still an open problem. A reliable prioritization (or ranking) of a gene is very important from a biological viewpoint, since it provides valuable information of a putative specific activity during the cell cycle. The smaller its rank position, the more likely it is a "true" CCR gene. This is particularly important, for example, in case of genes of unknown function, when gaining or losing a function after cell transformation due to cancer or other diseases or when comparing gene properties conservation across species, just to cite a few. It is also worth of note that the situation is similar to that of searching a term on Google: one wants the relevant pages to be presented at top positions. However, it is not clear yet which are the relevant features needed to fully characterize CCR genes and, on this basis, set up an efficient algorithm. One key issue is gene expression time-profiles susceptibility to many sources of signal distortions, in addition to measurement noise. The most important impacting factor is the technical artifact generated by cell synchronization methods that fail to maintain cell division synchrony, often shortly after the first cycle. Consequently, the expression time-profile of the second cycle is significantly different from the first one. From a computational perspective, a surprising outcome of this technical problem, is the poor consistency among ranked lists of CCR genes obtained from the same methodology applied to different datasets (Haase and Wittenberg, 2014). It suffices to mention the case of a 15% overlap among three datasets (reported in reference de Lichtenberg et al., 2005), to understand how disappointing this problem is. Although advances are being achieved, a clear *consensus* on the best method is still lacking (Doherty and Kay, 2010). Therefore, this subject is worthy of further investigation. Moreover, to motivate even more research on this issue, it is worth recalling the tight relationships between yeast and cancer cells in terms of the molecular machinery in charge of controlling cell cycle progression (Pray and Hartwell's, 2008).

There is a number of different methodologies for prioritizing (or ranking) cycling transcripts from gene expression time-profiles of biological processes like the mitotic cell-cycle, the metabolic cycle or the circadian rhythm, for example. Computational methods

* Corresponding author at: Department of Computer, Control and Management Engineering "A. Ruberti", Sapienza University of Rome, Italy.
E-mail addresses: lorenzo.farina@uniroma1.it (L. Farina), paola.paci@iasi.cnr.it (P. Paci).

https://doi.org/10.1016/j.jtbi.2018.09.025
0022-5193/

Table 1
Description of the available *Saccharomyces cerevisiae* mitotic cell cycle gene expression time-profile datasets used in this paper (part 1 of 3).

	$alpha_{SPE}$	$alpha_{GRA}$	$alpha30_{PRA}$	$alpha38_{PRA}$
Duration	119 min	200 min	120 min	120 min
Sampling	7 min	5 min	5 min	5 min
Cycles	2	3	2	2
Genes	6145	6378	4774	5006
Duplication time	62.3 min	65.7 min	64.0 min	64.0 min
Missing data	yes	no	yes	yes
Duplicates	no	no	no	no
Outliers	28.8%	37.0%	41.8%	56.9%
Skewness	13.1%	9.6%	21.9%	35.0%
Reference	Spellman et al. (1998)	Granovskaia et al. (2010)	Pramila et al. (2006)	Pramila et al. (2006)

can be roughly divided into two large groups: algorithms working in the time domain through some kind of "pattern matching" to a pre-defined time function chosen by the user (*e.g.* a sinusoid), and those working in the frequency domain through some kind of signal decomposition (*e.g.* Fourier series). See Doherty and Kay (2010) for a recent review.

Here, we will use as a source of information only gene expression time-profiles data, (*i.e.* we will not consider additional features like protein structures or binding motifs). Indeed, our aim is not to provide a general-purpose method to reveal underlying periodicities in time-series (as in Deckard et al., 2013), but to obtain a simple and fast computational tool able to prioritize genes according to their capacity to be regulated by the cell cycle, using budding yeast as a testbed. Our approach will mainly draw from the gene expression feature-based methodology proposed by de Lichtenberg et al. (2005) that will be called throughout the paper, the "*DL* algorithm". This methodology takes into account a combination of two features characterizing expression time-profiles of CCR genes: amplitude (which they called *regulation*) quantified by its standard deviation (SD), and cyclicity (which they called *periodicity*) quantified by a frequency-based score (called Fourier score) proposed by Spellman et al. (1998). In what follows, we will compare the performance of our algorithm (called *PERLA*) with the *DL* algorithm only, because the latter has been proved to outperforms more than 20 algorithms (Gauthier et al., 2008). Moreover, we will make use of the same benchmarks as in de Lichtenberg et al. (2005), so to make the comparison fair, and propose some new ones.

This paper supports the underlying rationale of the de Lichtenberg et al. method (de Lichtenberg et al., 2005) that provides scores for relevant features and combine them into a single one. This approach - in our opinion - has been the key to success for effective CCR gene prioritization. For this reason, we will further pursue this key idea and suggest a methodology called *PERLA* (PEriodicity, Regulation and LAg-autocorrelation) that introduces a new combined set of features aiming to significantly improve performances across an extended set of recent new experimental data of the budding yeast mitotic cell cycle.

1.1. Budding yeast cell cycle gene expression time-profiles datasets

The experimental gene expression time-profiles datasets used by de Lichtenberg et al. (2005) were obtained from three independent experiments performed on budding yeast (*Saccharomyces cerevisiae*). Two of them has been taken from the work of Spellman et al. (1998), and the third one from the work of Cho (1998) and re-normalized by de Lichtenberg et al. (2005). InTables 1–3 relevant information on all datasets used in this paper is summarized. Experiments whose name contain the term "alpha" are synchronized using the "alpha factor" protocol, those containing the term "cdc15" or "cdc28" are synchronized using the "temperature sensitive mutant" protocol and those containing the term "elu" are synchronized using the "elutriation" protocol (see Banfalvi, 2017 for a complete description of synchronization techniques). Finally, the subscripts refer to the first author of the corresponding paper. The following experiments are technical replicates: $alpha30_{PRA}$ and $alpha38_{PRA}$, $alphaR1_{ESE}$ and $alphaR2_{ESE}$, $eluR1_{ORL}$ and $eluR2_{ORL}$. The corresponding scores were averaged and denoted by, respectively, $alpha3038_{PRA}$, $alpha_{ESE}$ and elu_{ORL}. We pre-processed data using the following rules: *(i)* we allowed at most 10% of missing points for each gene expression time-series which were replaced by the corresponding points of the most correlated profile, *(ii)* duplicates data were averaged.

It is worth mentioning that, for the budding yeast mitotic cell cycle, currently, a total of 12 datasets (actually 9 if we do not consider technical replicates) produced by 6 independent groups using 4 different types of cell synchronization, are available. This is a rare and invaluable situation in molecular biology, where the lack of independent experiments at high time sampling rates are well known to greatly hinder data analysts from developing algorithms on dynamical systems. At the time of the de Lichtenberg et al. paper (de Lichtenberg et al., 2005), only a total of 3 experiments from 2 independent groups using 3 different types of synchronization were available. The current accessibility of 9 datasets (3 "old" and 6 "new") of better quality than the previous ones, provides the perfect testbed for re-considering the CCR gene prioritization problem on la large variety of experimental conditions and the optimal framework for robust comparative evaluation of the proposed method, *PERLA* precisely.

As previously stated, the most interesting issue arising when studying CCR genes prioritization, is the surprisingly weak consistency among ranked lists obtained from different methodologies on the same data, *i.e.* the small overlap among the top ranking genes. A common explanation is that current methods are not "smart" enough. However, this is only a part of the story, since multiple biological factors could contribute to this poor overlap. For example, some transcripts may be regulated only in specific conditions that are not consistent among laboratories and signal characteristics may be very sensitive to experimental factors like strain backgrounds, synchrony procedures, growth medium, or even microarray/NGS platform and raw data normalization procedures (Doherty and Kay, 2010; Haase and Wittenberg, 2014). Nevertheless, the proposed *PERLA* algorithm provides a substantial increase of the overlap among different experiments, thus proving that a computational effort in this direction is worth doing. Moreover, in the last paragraph we will briefly discuss the somewhat "ambiguous" matter regarding the "number" of CCR-specific genes. In fact, such number, as reported in the literature, varies from 416 to 1270! (Haase and Wittenberg, 2014). At the end of the day, which genes are "truly" CCR? The actual situation is that there is no single, clear-cut set simply because there is no universally agreed definition of what CCR exactly means (Haase and Witten-