Bioinformatics Advance Access published July 4, 2012

BasyLiCA: a tool for automatic processing of a Bacterial Live Cell Array

Leslie Aïchaoui [1], Matthieu Jules [2,3], Ludovic Le Chat [2,3], Stéphane Aymerich [2,3], Vincent Fromion [1] and Anne Goelzer [1]

[1]INRA, UR1077 Unité Mathématique Informatique et Génome, F-78350 Jouy en Josas, France
[2]INRA, UMR1319 Microbiologie de l'Alimentation au service de la Santé, F-78350 Jouy en Josas, France
[3]AgroParisTech, UMR1319 Microbiologie de l'Alimentation au service de la Santé, F-78350 Jouy en Josas, France

Associate Editor: Dr. Olga Troyanskaya

ABSTRACT

Summary: Live Cell Array (LCA) technology allows the acquisition of high-resolution time-course profiles of bacterial gene expression by the systematic assessment of fluorescence in living cells carrying either transcriptional or translational fluorescent protein fusion. However, the direct estimation of promoter activities by time-dependent derivation of the fluorescence datasets generates high levels of noise. Here, we present BasyLiCA, a user-friendly open-source interface and database dedicated to the automatic storage and standardised treatment of LCA data. Data quality reports are generated automatically. Growth rates and promoter activities are calculated by tunable discrete Kalman filters that can be set to incorporate data from biological replicates, significantly reducing the impact of noise measurement in activity estimations.

Availability: The BasyLiCA software and the related documentation are available at http://genome.jouy.inra.fr/basylica.

Contact: anne.goelzer@jouy.inra.fr.

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

1 INTRODUCTION

Despite recent developments in transcriptomic technologies, direct RNA sequencing and tiling array approaches require tedious sample preparations. For this reason, they are not suitable for transcriptional high-resolution time-courses that require one microarray and/or mRNA extraction per timepoint. An alternative approach is to measure the transcriptional activity of promoters using reporter genes. In the last decade, fluorescent proteins, in particular Green Fluorescent Protein (GFP), have become widely used (Giepmans *et al.*, 2006). Major advances have come from the characterisation and development of fast-folding GFP monomer variants (Griffin *et al.*, 1998). Fluorescent reporters allow the high-temporal-resolution measurement of promoter activity in living cells (Ronen *et al.*, 2002). Live Cell Array (LCA) technology involves the generation of a large collection of strains that harbour transcriptional fusions with fast-folding fluorescent proteins and monitoring their accumulation under the appropriate conditions. The promoter activity profiles of up to 96 individual *gfp* fusions in cells grown in microtiter plates can be obtained at once at very high resolution by determining the difference in fluorescence levels at successive timepoints. Promoter activation or deactivation can be easily detected by an increase or a decrease in the fluorescence accumulation rate. This high-throughput technology was proven to be an accurate and versatile method of determining gene expression in 2000 *Escherichia coli* promoters subjected to a glucose-lactose diauxic shift experiment (Zaslaver *et al.*, 2006).

We previously described the development of the plasmid pBaSysBioII, constructed within the framework of the EU-funded BaSysBio systems biology program (http://www.basysbio.eu/), for use in the LCA analysis of gene expression in *Bacillus subtilis* (Botella *et al.*, 2010). However, no tools have been developed to facilitate the analysis of the quickly growing bacterial LCA datasets. Here, we report the development of a user-friendly software, BasyLiCA, for the storage and analysis of LCA data. BasyLiCA is dedicated to wet lab biologists for the analysis of large amounts of LCA data in microplates. As a proof of concept, we produced and analysed LCA data for several promoters using *B. subtilis* as a model bacterium and the newly developed pBaSysBioII plasmid (see Supplementary information).

2 BASYLICA DESCRIPTION

BasyLiCA is an open-source software for the automatic and systematic management and treatment of LCA datasets. The software is compatible with Windows XP, Vista and 7 and published under the GNU licence. BasyLiCA is composed of a database, a web interface and a data analysis module, which is dedicated to the estimation of promoter activities and developed in the standard open-source R language (see Figure S2 and Supplementary information for details).

Database: The BasyLiCA database was developed in MySQL and is composed of seven tables (see Figure S3 in the Supplementary information) describing all of the parameters of LCA experiments: plate and strain characteristics, well composition, injection information, measured and treated data. The privacy of data and of strain characteristics can be easily managed via a dedicated administrator

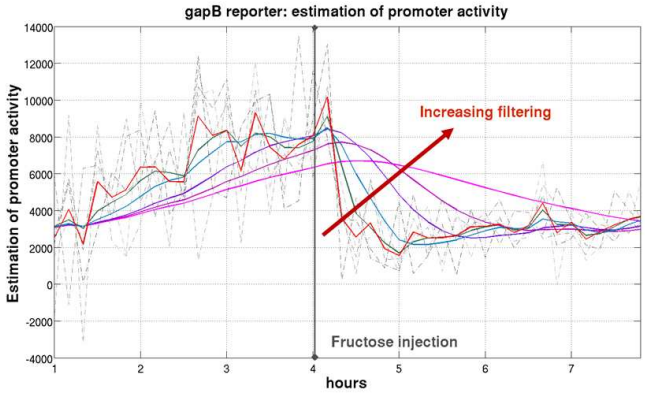

Fig. 1. Promoter activity of the *gapB* gene upon the addition of fructose (at t=4 h) on malate grown cells. Promoter activity was estimated incorporating or not LCA data from five biological replicates obtained with a sampling time of 10 minutes. Grey curves correspond to the promoter activity of each replicate independently computed by raw differentiation of the fluorescence over time divided by the OD. Color curves correspond to several estimations of the promoter activity calculated by Type I Kalman filter and using the five replicates from the smallest (in red) to the largest (in magenta) combination of filtering parameters.

module to allow access to the owner, the owner and colleagues or the entire world.

Interface: The web interface is implemented in php/html and can be used either locally or in web server mode through the Wampserver provided with BasyLiCA. The user-friendly interface allows (*a*) the automatic insertion of LCA measurements; (*b*) the manual or semi-automatic insertion of the characteristics of wells, strains, and injection; (*c*) the administration and the management of the database as a simple user or as an administrator; and (*d*) data treatment.

Data treatment: LCA data are pre-treated to evaluate the data quality. A report is automatically generated to help the user select the relevant wells for the estimation of the promoter activity. The data associated to the selected wells are then treated by well-established algorithms for filtering in industrial engineering, named discrete Kalman filters. Essentially, the discrete Kalman filter uses a series of measurements observed over time, containing noise and uncertainties, to produce estimates of unknown variables (promoter activities in the case of BasyLiCA). Theoretically, such estimations (using biological replicates) tend to be more precise than those solely based on a single measurement. Moreover, the data dynamics is free of *a priori* fitting functions (e.g. polynom, logistic...). The discrete Kalman filter only requires the definition of a dynamical model describing the theoretical behavior over time of LCA data:

$$\begin{cases} \mathrm{OD}(k+1) = (1+\mu(k)\delta_t)\mathrm{OD}(k) + b_o(k) \\ \mathrm{Fluo}(k+1) = \mathrm{Fluo}(k) + a(k)\mathrm{OD}(k)\delta_t + b_f(k) \end{cases}$$

where δ_t is the sampling time, $\mu(k)$ and $a(k)$ correspond to the growth rate and the promoter activity at time k, respectively, and $b_o(k)$ and $b_f(k)$ are the noise in the optical density (OD) and fluorescence (Fluo) measurements, respectively. All variables of the model are time-dependent. In BasyLiCA, two types of discrete Kalman filters are implemented, including (Type I) or omitting (Type II) data from biological replicates, and are based on the robust numerical algorithm of Verhaegen and Van Dooren (1986). The Type II discrete Kalman filter is applied to each well and estimates OD, fluorescence, growth rate and promoter activity. For the Type I filter, the optical density and growth rate are first estimated for each well of the plate. Then the fluorescence and the promoter activity are estimated by assuming a common promoter activity for all replicates. The Kalman filters can be adjusted by only two parameters representing the data quality level and the smoothing degree. These parameters are easily tunable by sliders in BasyLiCA. The results are stored in the database and in CSV and PDF files. Bacteria can also have a low level of auto-fluorescence, in which case additional post-treatment is required when computing the promoter activities (see Supplementary information).

3 APPLICATION

Promoter activities can be directly computed by raw differentiation of the fluorescence curve over time divided by the OD curve. However, the promoter activities obtained in this way are highly noisy (see Figure 1). By contrast, the discrete Kalman filter allows the user to set a trade-off between the estimation of the dynamics of the promoter activity and the level of noise filtering (colored curves). Furthermore, both the dynamics of the filter and the level of the noise filtering are improved if replicates are included in the discrete Kalman filter compared to the Type II (without replicates) Kalman filter. In Signal Processing, the decrease of the sampling time theoretically improves the estimations of promoter activity dynamics. However, lowering the sampling time to 1 minute had not the expected impact. Actually, technical noise increased, which requires high level of noise filtering (see Supplementary information). Consequently, a sampling time of 5 or 10 minutes should be more suitable for the current LCA technology.

In conclusion, BasyLiCA is a convenient and user-friendly piece of software dedicated to high-throughout LCA growing datasets in Systems Biology. It will help wet lab biologists as well as modellers to capture the dynamics of promoter activity time-courses.

Funding: European BaSysBio project (LSHG-CT-2006-037469) and European Basynthec project (n°FP7-244093).

REFERENCES

Botella, E., Fogg, M., Jules, M., Piersma, S., Doherty, G., Hansen, A., Denham, E. L., Le, C. L., Veiga, P., Bailey, K., Lewis, P. J., van Dijl, J. M., Aymerich, S., Wilkinson, A. J., and Devine, K. M. (2010). pBaSysBioII: an integrative plasmid generating *gfp* transcriptional fusions for high-throughput analysis of gene expression in *Bacillus subtilis*. *Microbiology*, **156**, 1600–1608.

Giepmans, B. N., Adams, S. R., Ellisman, M. H., and Tsien, R. Y. (2006). The fluorescent toolbox for assessing protein location and function. *Science*, **312**, 217–224.

Griffin, B. A., Adams, S. R., and Tsien, R. Y. (1998). Specific covalent labeling of recombinant protein molecules inside live cells. *Science*, **281**, 269–272.

Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A*, **99**, 10555–10560.

Verhaegen, M. and Van Dooren, P. (1986). Numerical aspects of different kalman filter implementations. *IEEE Transactions on Automatic Control*, **AC-31**(10), 907–917.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G., and Alon, U. (2006). A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods*, **3**, 623–628.