Maturation models of fluorescent proteins are necessary for unbiased estimates of promoter activity

Antrea Pavlou, Eugenio Cinquemani, Johannes Geiselmann, Hidde de Jong

Maturation models of fluorescent proteins are necessary for unbiased estimates of promoter activity

Antrea Pavlou[1,2], Eugenio Cinquemani[1,2], Johannes Geiselmann[1,2*+], and Hidde de Jong[1,2*+]

[1] *Univ. Grenoble Alpes, Inria, Grenoble, France.*
[2] *Univ. Grenoble Alpes, CNRS, LIPhy, Grenoble, France*
[*] *these authors contributed equally to this work*
[+] *corresponding authors*

Running title: Maturation models of reporter proteins

Abstract

Fluorescent proteins (FPs) are a powerful tool to quantitatively monitor gene expression. The dynamics of a promoter and its regulation can be inferred from fluorescence data. The interpretation of fluorescent data, however, is strongly dependent on the maturation of FPs since different proteins mature in a distinct way. We propose a novel approach for analyzing fluorescent reporter data by incorporating maturation dynamics in the reconstruction of promoter activities. Our approach consists in developing and calibrating mechanistic maturation models for distinct FPs. These models are then used alongside a Bayesian approach to estimate promoter activities from fluorescence data. We demonstrate by means of targeted experiments in *Escherichia coli* that our approach provides robust estimates, and that accounting for maturation is, in many cases, essential for the interpretation of gene expression data.

Statement of significance

Fluorescent proteins have been widely used for the understanding of gene expression dynamics in both prokaryotes and eukaryotes. The interpretation of fluorescence data, however, is dependent on the maturation of reporter proteins. Previous work has shown that the maturation dynamics of many fluorescent proteins is complex. We demonstrate that maturation can indeed introduce biases in the analysis of gene expression data. We therefore combine mathematical modelling and statistical inference to propose a novel approach for the analysis of gene expression data that is capable of reconstructing promoter activities in a robust manner, correcting for the maturation dynamics of specific fluorescent proteins. The approach has been validated by means of an experimental system in *Escherichia coli*.

Introduction

Since the discovery of GFP in 1992 [1], hundreds of fluorescent proteins (FPs) covering the visible light spectrum have been developed and become essential for the visualization and quantification of biological phenomena in living cells [2]. FPs notably allow the quantitative investigation of gene expression in a dynamical context [3]. Accordingly, many computational methods have been developed in order to quantitatively reconstruct the promoter dynamics from FP timelapse measurements [4, 5, 6, 7, 8, 9, 10].

An observed fluorescence signal is influenced by the distinct physical characteristics of a FP, notably maturation, whose mechanisms and kinetics vary significantly across proteins [11]. The maturation kinetics may decouple FP production and fluorescence emission, thus introducing significant biases in data analysis and interpretation. No method currently exists that incorporates this information in the reconstruction of promoter activities.

In this paper, we use a combination of mathematical modeling, experimental calibration and statistical inference to integrate maturation dynamics in the reconstruction of promoter activities from fluorescence data. We focus on the two most used FPs, green fluorescent protein (GFP) and red fluorescent protein (RFP) which have very different maturation dynamics [11]. In particular, we developed appropriate ODE models taking into account what is currently known about the maturation mechanism of each FP and calibrated the models using experimental data. We also developed a Bayesian inference approach to robustly reconstruct promoter activities from timelapse fluorescence data. To validate our approach, we constructed an experimental system in *Escherichia coli* where the two reporters are upstream of the same constitutive promoter. We show that, with correction for maturation, the inference procedure yields the same (normalized) promoter activities for the two FPs. Without correction, however, the promoter activities diverge. In particular, for the slowly-maturing RFP, the promoter activities are underestimated and have a delayed dynamics. This demonstrates that accounting for maturation is crucial for the correct analysis and interpretation of fluorescence data.

Our principled and practically applicable approach can be used for the robust reconstruction of promoter activities in multiple growth conditions and for both prokaryote and eukaryote reporter systems.

Results

Mechanistic maturation models

To correctly assess maturation effects in dynamic conditions, we first defined mechanistic models for GFP and RFP, in particular the chosen variants GFPmut2 and mScarlet-I, taking into account maturation mechanisms reported in the literature. For GFPs in general, an immature colorless species is transformed into a mature green species [12]. Maturation of mScarlet-I and other RFP variants is more complex: a colorless species is

transformed into a mature red species via the formation of a blue intermediate absorbing at around 400 nm [13, 14]. The different maturation mechanisms of GFP and RFP lead to distinct maturation kinetics [11], and correspondingly, to different maturation models. Whilst a GFP model has a single maturation step, an RFP model requires two steps in order to account for the blue intermediate (Fig. 1A-B).

The models are defined on the population-level: they describe the total quantity of protein species (red, green, blue, colorless) in a growing population of cells. The quantities of protein are assumed proportional to the measured fluorescence signals, reported in relative fluorescence units (RFU). The proportionality factors are different for green, red, and blue FPs, due to differences in brightness of the FPs, gains of the photomultiplier tubes, *etc.* We therefore introduced $\mathrm{RFU}_{\mathrm{green}}$, $\mathrm{RFU}_{\mathrm{red}}$, and $\mathrm{RFU}_{\mathrm{blue}}$ to distinguish between measurements of green, red, and blue fluorescence. Biomass is quantified by absorbance (Abs), assumed proportional to the volume of the bacterial population. Little is known about the reversibility of the maturation reactions of RFP, so we have constructed and compared several variants of the RFP model that include or exclude backflow reactions (Fig. 1A-B and Text S1). Statistical model selection showed that the best model for each protein is the most parsimonious one, without backflows (*Materials and methods* and Text S1).

The GFP model is composed of two ordinary differential equations describing the dynamics of the quantity of immature proteins $Im(t)$ [$\mathrm{RFU}_{\mathrm{green}}$] and the quantity of mature green proteins $M(t)$ [$\mathrm{RFU}_{\mathrm{green}}$]. Note that expressing the quantities of mature and immature proteins in the same units allows for their direct comparison. The rate of production of immature proteins is given by $\alpha(t) \cdot V(t)$, where $\alpha(t)$ [$\mathrm{RFU}_{\mathrm{green}}$ min^{-1} Abs^{-1}] is the specific production rate, per unit population volume, and $V(t)$ [Abs] the volume of the growing population. The specific production rate is also called promoter activity in the literature, where it is assumed that the dynamics of the intermediate mRNA species can be ignored [7]. The conversion of immature to mature protein occurs at a rate proportional to the quantity of immature protein with constant k_m [min^{-1}], and all proteins are degraded at the same rate with constant γ [min^{-1}]:

$$\frac{d}{dt}Im(t) = \alpha(t) \cdot V(t) - (\gamma + k_m) \cdot Im(t), \tag{1}$$

$$\frac{d}{dt}M(t) = k_m \cdot Im(t) - \gamma \cdot M(t). \tag{2}$$

The RFP model describes the dynamics of the quantities of immature proteins $Im(t)$, intermediate blue proteins $Hm(t)$, and mature red proteins $M(t)$, with appropriate rate constants for inter-species conversion k_{hm}, k_m.

$$\frac{d}{dt}Im(t) = \alpha(t) \cdot V(t) - (\gamma + k_{hm}) \cdot Im(t), \tag{3}$$

$$\frac{d}{dt}Hm(t) = k_{hm} \cdot Im(t) - (\gamma + k_m) \cdot Hm(t), \tag{4}$$

$$\frac{d}{dt}M(t) = k_m \cdot Hm(t) - \gamma \cdot M(t). \tag{5}$$

For ease of comparison, all quantities are expressed in units $\mathrm{RFU_{red}}$, even the intermediate blue proteins. In order to relate the variable Hm to the observed blue fluorescence, we introduce a rescaling factor z [$\mathrm{RFU_{red}}$ $\mathrm{RFU_{blue}^{-1}}$] and set $Hm = z\,\overline{Hm}$, with $\overline{Hm}$ expressed in units $\mathrm{RFU_{blue}}$.

To estimate the kinetic maturation parameters of the two models, we conducted growth-arrest experiments where we quantified the fluorescence intensity of each species after adding Chloramphenicol (Cm) to stop translation. As a consequence, $\alpha(t) = 0$ and every increase in fluorescence after adding antibiotics is due to maturation. Whilst extensive growth-arrest experiments have been conducted by Balleza *et al.* [11] to quantify the kinetics of mature FPs, we did our own experiments to also track and quantify the blue intermediate in RFP maturation (*Materials and methods*). The data from these experiments were also used to estimate the degradation constants.

The maturation curves for RFP are shown in Fig. 1C along with the dynamics of its blue intermediate. The estimation of the model parameters k_{hm} and k_h from the two curves (*Materials and methods*, Text S1), gives rise to an excellent fit ($R^2_{blue} = 0.91$, $R^2_{red} = 0.99$). In parallel, the same experiment was conducted with a strain expressing GFP. The green fluorescence curve was used to calibrate the parameter k_m in the GFP model (Fig. 1D, $R^2_{green} = 0.97$). The values of the parameters thus obtained demonstrate the difference in maturation time for the two proteins: the GFP model parameter indicates fast maturation (9 min), whereas for the RFP model we find significantly slower maturation for the rate-limiting first step (35 min) (Text S1). These maturation times agree quite well with those reported in previous work [11]: 6 min for GFP and 26 min for RFP.

In conclusion, we thus obtained kinetic models, tailored for each FP and its maturation mechanisms, and calibrated these models by means of targeted experiments.

Bayesian approach for the reconstruction of promoter activities from fluorescence data

Leveraging the models presented above, we developed an approach to reconstruct an unknown promoter activity profile $\alpha(t)$ from measurements of fluorescent reporter abundance $M(t)$. We assume that K time-sampled, noisy measurements $\tilde{M}_k = M(t_k) + e_k$ are available, where t_k are the measurement times and e_k random measurement errors with assigned variance (for $k = 1, \ldots, K$). $V(t)$ can be determined directly from absorbance data. Therefore, we first address the problem of estimating the time profile of the whole term $A(t) = \alpha(t) \cdot V(t)$ (the aggregated synthesis rate of immature proteins over the whole population), and then deduce the promoter activity $\alpha(t)$, which is the biologically relevant quantity, in a post-processing step.

Reconstructing the input (here $A(t)$) of an ODE system from sampled output measurements (here $\tilde{M}_k$) is a nontrivial inverse problem [15]. Pure data fitting results in

irregular and highly uncertain estimates of the profile $A(t)$ due to measurement noise and data sparsity. A robust estimation of $A(t)$ at any time t within the measurement period can be obtained by Bayesian regularization [16]. According to the Bayesian paradigm, a prior probability distribution, say $f(A)$, is introduced to express *a-priori* belief in the unknown profile $A(t)$. Given measurements $\tilde{M}$, the posterior distribution of A, $f(A|\tilde{M})$, conveys the updated knowledge on A. At any time t, estimates of $A(t)$ follow from this posterior. In particular, the mean of this posterior,

$$\hat{A}(t) = \mathbb{E}[A(t)|\tilde{M}_1, \ldots, \tilde{M}_K], \tag{6}$$

provides estimates $\hat{A}(t)$ that statistically minimize the estimation error [17, 18]. The posterior is formally obtained via the Bayes rule $f(A|\tilde{M}) \simeq f(\tilde{M}|A) \cdot f(A)$. The likelihood $f(\tilde{M}|A)$ is determined by the solution of Eq. 1-2 or Eq. 3-5 as a function of A and the measurement error statistics. In the context of regularization, a suitable choice of $f(A)$ plays the role of a regularization term that balances the data fitting expressed by the likelihood with the expected properties of A.

Calculating the posterior $f(A|\tilde{M})$ via the likelihood $f(\tilde{M}|A)$ is generally cumbersome. Here, we exploit linearity of the maturation models and a choice of the prior $f(A)$ that make the calculations to obtain estimates Eq. 6 and their error variance very efficient.

Borrowing from Rasmussen *et al.* [19], we let $f(A)$ be a Gaussian process prior. That is, we let $A(t)$ follow the laws of a (zero-mean) Gaussian stochastic process, for which the autocorrelation function (referred to as the kernel in Gaussian process learning) entirely defines the process properties. In absence of specific information on the statistics of A at different times t, we assume the process to be stationary. We choose an exponential kernel, which is characterized by two parameters $\lambda > 0$ and $\theta > 0$ (see *Materials and methods* for the mathematical details). Smaller values of λ (respectively, θ) assign higher probability to small (respectively, slower) fluctuations of $A(t)$.

For the purpose of estimating A from the data, λ and θ play the role of regularization parameters: small values of λ and θ enforce stronger regularity of the solution, but the correspondingly smooth predictions of $M(t)$ may not match fast transitions in the observed fluorescence time-series. Conversely, large values of λ and θ lead to perfect data interpolation at the price of overly irregular estimates of $A(t)$. In general, appropriate values for these parameters are not available *a priori*. However, in our case, optimal values $\hat{\theta}$ and $\hat{\lambda}$ for θ and λ can be determined directly from the data by maximizing the (marginal) likelihood $f_{\lambda,\theta}(\tilde{M})$ using fast numerical procedures (*Materials and methods*).

For regularization parameters fixed as above, the estimation task is now to calculate Eq. 6 for the chosen prior on A. In our approach, rather than constructing estimates in terms of the expression of the kernel [19], we profit from the stochastic differential equation that characterizes process A. When combined with the maturation model (Eqs 1-2 or Eqs 3-5), this description of $A(t)$ gives rise to a linear stochastic differential equation system for which Eq. 6 can be efficiently computed via so-called Kalman filtering/smoothing ([17, 20] and *Materials and methods*). The method is exact for Gaussian measurement

error, and it remains viable for moderate deviations from Gaussianity since it always yields the best estimator among linear functions of the data [17].

In summary, our method for the estimation of the promoter activity profile $\alpha(t)$ works as follows (Fig. 2). Given time-course data from a reporter gene experiment, we first estimate optimal regularization parameters $\hat{\lambda}$ and $\hat{\theta}$, then use these to calculate the optimal estimate $\hat{A}_{\hat{\lambda},\hat{\theta}}(t)$ at all times t of interest as per Eq. 6. Finally, we define our estimate of $\alpha(t)$ as $\hat{\alpha}(t) = \hat{A}_{\hat{\lambda},\hat{\theta}}/\hat{V}(t)$, where $\hat{V}(t)$ is given by the absorbance data. The robust performance of the approach was verified by means of synthetic data (Fig. S2).

Protein-specific maturation models provide robust estimates of promoter activities

To investigate whether maturation must be taken into account for the correct reconstruction of promoter activities, we used the Bayesian estimation method alongside the calibrated maturation models to interpret fluorescence data obtained from a controlled experimental setup.

Specifically, we designed an experimental system in *Escherichia coli* where a red and a green FP are under the influence of the same constitutive promoter, proC [21]. This makes it possible to directly and quantitatively compare the promoter activity profiles for each protein (*Materials and methods*). The two newly-constructed strains were used to perform kinetic experiments in batch, where the absorbance and red and green fluorescence were measured during growth in minimal medium with four different carbon sources: glucose, xylose, acetate and pyruvate (Figs S3-S10). Absorbance and fluorescence curves for glucose and xylose are shown in Fig. 3B-C. Each curve was then used to reconstruct the underlying promoter activity $\alpha(t)$ from the models schematized in Fig. 3A, following the approach in Fig. 2.

As can be seen in Fig. 3D-E, the promoter activities estimated from the green and red fluorescence data are very similar. For both glucose and glycerol, the activities are steady during exponential growth and drop to a negligible value when the carbon source is exhausted. For the chosen machine settings, the GFP and RFP promoter activities are not only similar in shape, but also in magnitude. In general, this will not be the case, because they are expressed in different units ($\mathrm{RFU_{green}}$ $\mathrm{min^{-1}}$ $\mathrm{Abs^{-1}}$ *vs.* $\mathrm{RFU_{red}}$ $\mathrm{min^{-1}}$ $\mathrm{Abs^{-1}}$). If the proportionality factors relating green and red fluorescence to GFP and RFP quantities were known, the promoter activities could be recomputed in the same scale. This is not usually the case, however, and the precise determination of the proportionality factors is a hard task requiring dedicated calibration experiments [22]. In order to compare the GFP and RFP promoter activities, we therefore normalize them, here with respect to the mean during steady-state exponential growth (between 200 and 350 min). The normalized promoter activities are shown in the inserts of Fig. 3D-E.

If maturation affects the inference of promoter activities, one would expect the latter to change when maturation is ignored. In order to test this, we inferred $\alpha(t)$ from the

fluorescence data by means of the model

$$\frac{d}{dt}M(t) = \alpha(t) \cdot V(t) - (\gamma + k_m) \cdot M(t), \tag{7}$$

in which the maturation step has been omitted. The results of the estimation process are shown in Fig. 3F-G.

When comparing the promoter activities obtained with and without maturation correction, a first observation is that in the case of RFP their magnitude is underestimated when ignoring maturation. For GFP, however, which has a much shorter maturation time, this effect is not evident from the plots. In order to quantify the observed changes, we computed the R^2 value of pairs of RFP promoter activities with and without maturation correction. For growth on glucose and xylose, we found values of 0.77 and 0.66, respectively. In the case of GFP, these values are much higher, indicating that the promoter activities are much more similar: $R^2 = 0.99$ for both glucose and xylose.

A second observation is that, when ignoring maturation, the normalized promoter activities for GFP and RFP, which coincide after maturation correction, start to diverge. In particular, the dynamics of the RFP promoter activities are delayed, as shown in the inserts of Fig. 3F-G. We quantified these delays as the difference in time for the normalized activities to reach their half-maximal value. When correcting for maturation, this delay is negligible, on the order of minutes ($\tau_{glucose} = 3$ min and $\tau_{xylose} = 7$ min). When ignoring maturation, however, the delays are substantial: $\tau_{glucose} = 39$ min and $\tau_{xylose} = 53$ min.

Similar observations as above can be made for the other two carbon sources considered, acetate and pyruvate (Figs S7-S10). For example, the R^2 value of pairs of RFP promoter activities with and without maturation correction equals 0.75 in the case of pyruvate (0.99 for GFP). Together these results indicate that maturation correction is necessary to avoid the introduction of a bias in the reconstruction of promoter activities, especially for slow-maturing FPs.

In order to achieve maturation correction, we used a protein-specific two-step model for RFP (Eqs 3-5), consistent with what is known about the biochemical reactions underlying maturation of this protein (Fig. 1B). How would the results change if we had instead opted for a simpler, one-step model, like that used for GFP? We fitted Eqs 1-2 to the red fluorescence data from the calibration experiment (Fig. 1C) and inferred $\alpha(t)$ from the red fluorescence data in the validation experiment (Fig. 2A-B). As shown in Fig. S11, the results are very similar to those obtained with the two-step model, even when using a published instead of estimated maturation constant [11]. This means that, at least in the case considered here, the correction for maturation *per se* is more important than the specific maturation model used. One reason that the one-step model yields a good approximation in this case is that the time-scale of the dynamics of the two-step model is essentially determined by the first step, which is much slower than the second step (Text S1).

Discussion

Fluorescent proteins have become a powerful tool for monitoring gene expression in living cells. The relation between the fluorescence intensity emitted by the cells and the amount of protein is not straightforward, because it depends on specific properties of the FPs. In this work, we investigated the effect of a key property, maturation, which was shown to vary significantly across FPs [11]. Ignoring maturation when reconstructing promoter activity and other measures of gene expression may introduce a bias due to the dynamic decoupling of fluorescence intensity and FP production. We developed a Bayesian inference approach, based on protein-specific models of maturation and Kalman filtering/smoothing algorithms, to correct for this bias. Moreover, we validated the approach using a tailored experiment system, consisting of *E. coli* strains carrying identical plasmid systems with different fluorescent reporter genes.

The maturation models considered concern the case of populations of microbial cells, typically growing in the wells of a microplate. The motivation for this focus is that thermostated microplate readers have made automated and multiplexed reporter gene assays on the population level accessible to almost any microbiology laboratory today [7, 8, 10]. It confers a broad relevance to the question of how to correctly interpret the resulting data. The same question can be posed, of course, for single-cell reporter gene assays carried out by means of time-lapse fluorescence microscopy [23]. Generalization of the approach to single-cell experiments requires a reformulation of the models of Eqs 1-5.

The maturation models were calibrated by means of dedicated experiments, resulting in estimates of the kinetic rate constants (*Materials and methods* and Text S1). An identifiability analysis showed that parameter uncertainty is very low in the case of the models without backflow retained for the analysis (Fig. S14). This has motivated the use of point estimates for the maturation and degradation parameters in the Bayesian inference procedure reconstructing the promoter activities. In order to make sure that the effect of parameter uncertainty is indeed negligible, we repeated the analysis of the data in Fig. 3B-C for alternative parameter values within the uncertainty interval. The results are not significantly different from those reported for the reference parameter values (Fig. S15), which ensures that parameter uncertainty does not affect conclusions about the importance of maturation correction.

The results reported in Fig. 3 and Figs S3-S10 provide a clear indication of the situations in which maturation correction is important. Ignoring maturation does not change much the promoter activities inferred from fluorescence data for FPs with short maturation times like GFPmut2 (Fig. 3F-G). In the case of FPs with long maturation times like mScarlet-I, however, it leads to an underestimation of promoter activities and delayed dynamics. While we used two-step maturation models for the analysis of RFP data, consistent with what is known about the maturation mechanisms, even a generic one-step model is able to avoid most of these problems in the case of the FPs considered here (Fig. S11). Maturation correction may generally not be sufficient for obtaining unbiased

estimates of promoter activity, since there may be other sources of bias beyond fluorophore maturation, such as coding bias and translation initiation efficiency. However, our results show that maturation correction is necessary for proteins with non-negligible maturation times, which is the case for many if not most FPs [11].

A well-known, perturbing factor in the use of fluorescent reporter gene assays is oxygenation of the microbial culture, due to the requirement of oxygen for FP maturation [7]. Growth to high biomass concentrations in standard batch cultures causes oxygen partial pressure pO_2 to drop to low levels [24]. The results presented in Fig. 3 and Figs S3-S10 were obtained in media with low substrate and therefore low biomass concentrations. We expect deviations from the inferred profiles when increasing the substrate concentration, which we verified by growing our reporter strains in a medium with 0.2% instead of 0.05% glucose (Fig. S12). In the case of mScarlet-I, a dip in promoter activity occurs after 400 min, when the absorbance exceeds 0.1. The promoter activity recovers after 500 min, when glucose is exhausted and the *E. coli* cells continue growth on acetate secreted during growth on glucose. The growth rate on acetate is much lower than the growth rate on glucose, which reduces the oxygen utilization rate of the culture and therefore partially restores pO_2 [24]. This decreased oxygen utilization liberates oxygen for mScarlet-I maturation and could explain the recovery of the promoter activity. No dip in promoter activity is observed for GFPmut2, which may be due to the lower oxygen requirements for maturation of this FP (a single oxygen molecule instead of two for mScarlet-I, [25, 26]).

In conclusion, we provide a powerful approach for the reconstruction of promoter activities from gene expression data by incorporating FP maturation in the analysis of fluorescence data. While validated on a bacterial model system, the approach is directly applicable to reporter proteins in other prokaryotes and in eukaryotes. The only prerequisite is the availability of a calibrated maturation model. The current work illustrates how to develop such models and calibrate them by means of targeted experiments. When this is not possible, our results indicate that, under certain conditions, a generic model combined with published maturation times may go long a long way in correcting for a maturation bias.

Materials and methods

Bacterial strains and plasmids

For all experiments, an *Escherichia coli* BW25113 strain was used that contains a deletion of the *fhuA* gene, making it resistant to phage contaminations (Δ*fhuA*). The *pEB2-mScarlet-I* reporter plasmid [11] was transformed into the above strain using $CaCl_2$. A corresponding *pEB2-GFPmut2* plasmid was constructed using Gibson assembly [27]. In both plasmids, transcription of the reporter gene is controlled by the same constitutive promoter, proC [21]. The plasmids have a pSC101 origin of replication, meaning they are low-copy and the number of copies does not vary over different growth phases [28]. The

above resulted in the strains BW25113-Δ*fhuA-pEB2-mScarlet-I* and BW25113-Δ*fhuA-pEB2-GFPmut2*, which were used for all experiments. The protein sequences of the two reporters used in this study were aligned with the original GFPmut2 and mScarlet-I sequences published in the literature (FPbase IDs [29]: QS3XQ and 6VVTK, respectively). The two GFP sequences were identical, whereas a single deletion (Δ2V) was detected in the RFP sequence used in comparison with the reference sequence. For primers and sequences, see Fig. S13 and File S2.

Experimental conditions

The two strains described above, along with a control strain without plasmids, were grown overnight at 37°C, with shaking at 200 rpm in MOPS minimal medium, supplemented with 0.2% glucose. Pre-cultures were inoculated at an OD600 of 0.001 in a 96-well microplate where each well contained MOPS minimal medium [30] supplemented with 0.2% glucose.

In the calibration experiments, bacteria were incubated at 37°C in a microplate reader (Tecan Infinite M200 Pro) until reaching an uncorrected absorbance of approximately 0.3. 300 μg/mL of Chloramphenicol (Cm) was then added in each well to stop protein translation. The absorbance, green, blue and red fluorescence intensity were tracked for another 10 hours after translation arrest. The residual growth after Cm addition was negligible (Fig. S1).

For the validation experiments, we used MOPS minimal medium supplemented with 0.05% of glucose, xylose, acetate or pyruvate. This time, the bacteria were incubated in a 96-well microplate without the addition of antibiotics at 37°C for 24 hours. Absorbance, red, green and blue fluorescence were measured. For all experiments, the background was corrected using fluorescence measurements from the strain without plasmids. Glass beads were added to improve oxygenation and the resulting outliers were filtered using WellInverter [31].

Parameter estimation and model selection

The parameters of our maturation models, which are ordinary differential equation systems in the variables Im, Hm and M (abundance of immature, half-mature and mature fluorescent reporter proteins), were estimated from the results of the calibration experiment described above. As explained in Text S1, the short-term dynamics of the fluorescence curves (tens of minutes) served for the estimation of the maturation constants as well as the scaling factor z, which relates blue and red fluorescence measurements by $Hm = z \cdot \overline{Hm}$. The long-term dynamics (tens of hours) allowed the estimation of

the degradation constant γ. The estimation procedure used the least-squared method implemented in Python.

Four variants of the two-step maturation model were considered, having zero, one or two backflows (Text S1). We used the Akaike Information Criterion (AIC) statistic to select the best model among the four candidates, which resulted in the model with no backflows. An identifiability analysis by means of a bootstrapping procedure showed that the parameters of this model are identifiable with low uncertainty (Fig. S14). The model was then used in the Kalman filtering/smoothing algorithm for the reconstruction of promoter activities. More details can be found in Text S1.

Implementation of promoter activity estimation via Kalman smoothing

Promoter activity is expressed in terms of a time-varying profile $A(t)$. Abundance of mature fluorescent reporter proteins, $M(t)$, depends on $A(t)$ *via* the maturation models of Eqs 1-2 and Eqs 3-5, written in the form $\dot{x}(t) = Bx(t) + CA(t)$, where $M(t)$ is an entry of vector $x(t)$ together with $Im(t)$ and (where applicable) $Hm(t)$. Matrices B and C are defined by the kinetic parameters occurring in the maturation models.

In order to estimate $A(t)$ from fluorescence measurements $\tilde{M}_k = M(t_k) + e_k$ taken at discrete times t_k and corrupted by (Gaussian) noise e_k ($k = 1, \ldots, K$), we relied on the use of a Gaussian process prior on $A(t)$ expressed in the form of a stochastic differential equation. Together with a maturation model of interest, this makes it possible to write an augmented system of linear stochastic differential equations and to apply Kalman filtering and smoothing [17] to find the statistically optimal, computationally efficient solution of the estimation problem.

We define the prior on $A(t)$ by assuming that $A(t)$ follows the laws of an Ornstein-Uhlenbeck process [19]. That is, for positive parameters λ and θ, A obeys the stochastic differential equation

$$dA(t) = -\theta \cdot A(t) \cdot dt + \lambda \cdot dW(t), \tag{8}$$

where $W(t)$ is the standard Wiener (white noise) process. Choosing $A(0)$ as a zero-mean Gaussian random variable with variance $\lambda^2/2\theta$ ensures process stationarity. The autocorrelation function of A (correlation between $A(t)$ and $A(s)$ for any two times t and s) is then the exponential

$$\frac{\lambda^2}{2\theta} \cdot \exp(-\theta \cdot |t - s|).$$

It follows from this expression that θ statistically characterizes the memory of the process (how fast $A(t)$ may change over time), while the ratio $\lambda^2/2\theta$ characterizes the magnitude (variance) of $A(t)$. For the estimation problem, λ and θ play the role of regularization parameters, *i.e.*, they ensure that reconstruction of $A(t)$ is robust to measurement noise and sparse sampling.

Assume that optimal values $\hat{\lambda}$, $\hat{\theta}$ of λ and θ have been determined (see below). Combining Eq. 8 with the maturation model of interest yields the augmented system

$$d\begin{bmatrix} x(t) \\ A(t) \end{bmatrix} = \begin{bmatrix} B & C \\ 0 & -\hat{\theta} \end{bmatrix} \cdot \begin{bmatrix} x(t) \\ A(t) \end{bmatrix} dt + \begin{bmatrix} 0 \\ \hat{\lambda} \end{bmatrix} dW(t), \tag{9}$$

$$\tilde{M}_k = \begin{bmatrix} D & 0 \end{bmatrix} \cdot \begin{bmatrix} x(t_k) \\ A(t_k) \end{bmatrix} + e_k, \tag{10}$$

for an appropriate 0-1 row vector D such that $D \cdot x(t_k) = M(t_k)$. Let $v(t_k)$ be the augmented state vector comprising $x(t_k)$ and $A(t_k)$. For a generic index j, the optimal Bayesian estimate of $v(t_k)$ from the data up to time t_j is the conditional expectation $\hat{v}^j(t_k) = \mathbb{E}[v(t_k)|\tilde{M}_1, \ldots, \tilde{M}_j]$. In particular, we are interested in the calculation of $\hat{v}^K(t_k)$ (optimal estimates of v from all the K measurements) at all times t_k, from which optimal estimates of A as well as of Im and (where applicable) Hm follow (they are all entries of v). This was obtained by the following two-sweep data processing algorithm.

In a first sweep (Kalman filtering), estimates $\hat{v}^k(t_k)$ along with one-step predictions $\hat{v}^{k-1}(t_k)$ are computed iteratively for k from 1 up to K, together with corresponding estimation error variance matrices $P^k(t_k)$ and $P^{k-1}(t_k)$. The iteration is initialized by definitions of $\hat{v}^0(t_1)$ and $P^0(t_1)$ corresponding to a noninformative (large variance) prior for $x(t_1)$ and the stationary statistics of A for $A(t_1)$. In a second sweep (Kalman smoothing), estimates $\hat{v}^K(t_k)$ exploiting both past and future measurements are computed at all times t_k by back-processing $\tilde{M}_k$ and the estimates from the first sweep in a time-reversed iteration (k from K down to 1). This second (often neglected) sweep provides a significant refinement of the estimates [32]. The filtering and smoothing formulas are standard [17, 18]. Their implementation can be found in the supplementary code (File S1).

Given several experimental replicates (datasets $\tilde{M}_1, \ldots, \tilde{M}_K$), we calculated promoter activity estimates $\hat{A}$ separately for every replicate. Confidence intervals on the estimates $\hat{A}$ were then calculated as twice the standard error of the mean of the estimates obtained over the different replicates. Confidence intervals for the estimates of Im, Hm and M were obtained in a similar manner.

Optimal choice of regularization parameters

An optimal choice of λ and θ can be determined by solving the maximum likelihood problem

$$(\hat{\lambda}, \hat{\theta}) = \arg\max_{(\lambda,\theta)} f_{\lambda,\theta}(\tilde{M}_1, \ldots, \tilde{M}_K), \tag{11}$$

where $f_{\lambda,\theta}(\,\cdot\,)$ is the probability density function for the observed data under the parameter-dependent prior in Eq. 8. In practice, for any value of (λ, θ), evaluation of this likelihood can be performed by the same filtering tools described above. Generalizing previous

definitions, let $\hat{v}_{\lambda,\theta}^{k-1}(t_k)$ and $P_{\lambda,\theta}^{k-1}(t_k)$ be the optimal one-step prediction of $v(t_k)$ and corresponding error variance matrix under generic values of λ and θ. The calculation of these quantities used the same filtering sweep described in the previous section. In the light of Eq. 10, one can then evaluate at any k the conditional Gaussian densities

$$f_{\lambda,\theta}(\tilde{M}_k|\tilde{M}_1,\ldots,\tilde{M}_{k-1}) = \frac{1}{\sqrt{2\pi\Lambda_k}}\exp\left[-\frac{1}{2}\cdot\frac{\left(\tilde{M}_k-\widehat{M}_k^{k-1}\right)^2}{\Lambda_k}\right],$$

where $\widehat{M}_k^{k-1} = \begin{bmatrix}D & 0\end{bmatrix}\cdot\hat{v}_{\lambda,\theta}^{k-1}(t_k)$ and $\Lambda_k = \begin{bmatrix}D & 0\end{bmatrix}\cdot P_{\lambda,\theta}^{k-1}(t_k)\cdot\begin{bmatrix}D & 0\end{bmatrix}^T + \sigma_k^2$. In turn, σ_k^2 is the variance of the measurement error e_k, as determined from the variance of $\tilde{M}_k$ across multiple experimental replicates. Thus, using Bayes' law, the likelihood in Eq. 11 can be evaluated in terms of the above densities as

$$f_{\lambda,\theta}(\tilde{M}_1,\ldots,\tilde{M}_K) = \prod_{k=1}^{K} f_{\lambda,\theta}(\tilde{M}_k|\tilde{M}_1,\ldots,\tilde{M}_{k-1}).$$

For numerical convenience, the optimization problem of Eq. 11 was rather solved by the minimization of the negative logarithm of the likelihood, using Python solver `minimize` of the `scipy.optimize` module.

Supplementary material

Supplementary Material is available online.

Acknowledgments

This study is partly funded by the ANR projects MAXIMIC and CTRL-AB (ANR-17-CE40-0024 and ANR-20-CE45-0014, https://www.anr.fr) and the Inria IPL CoSy (https://www.inria.fr). Funding was awarded to AP, EC, JG, HdJ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

All relevant data are within the manuscript and the Supplementary Material files.

Author contributions

Conceived and designed the experiments: AP, EC, JG, HdJ. Performed the experiments: AP. Analyzed the data: AP. Wrote the Kalman filtering/smoothing code: EC. Wrote the paper: AP, EC, JG, HdJ.

Declaration of interests

The authors declare no competing interests..

Supporting citations

References [33, 34] appear in the Supporting Material.

References

[1] Tsien, R. Y., 1998. The green fluorescent protein. *Annu Rev Biochem* 67:509–544.

[2] Day, R. N., and M. W. Davidson, 2009. The fluorescent protein palette: tools for cellular imaging. *Chem Soc Rev* 38:2887–2921.

[3] Specht, E. A., E. Braselmann, and A. E. Palmer, 2017. A critical and comparative review of fluorescent tools for live-cell imaging. *Annu Rev Physiol* 79:93–117.

[4] Leveau, J. H. J., and S. E. Lindow, 2001. Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *J Bacteriol* 183:6752–6762.

[5] Wang, X., B. Errede, and T. C. Elston, 2008. Mathematical analysis and quantification of fluorescent proteins as transcriptional reporters. *Biophys J* 94:2017–2026.

[6] Finkenstädt, B., E. A. Heron, M. Komorowski, K. Edwards, S. Tang, C. V. Harper, J. R. E. Davis, M. R. H. White, A. J. Millar, and D. A. Rand, 2008. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics* 24:2901–2907.

[7] de Jong, H., C. Ranquet, D. Ropers, C. Pinel, and J. Geiselmann, 2010. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Syst Biol* 4:55.

[8] Lichten, C., R. White, I. Clark, and P. Swain, 2014. Unmixing of fluorescence spectra to resolve quantitative time-series measurements of gene expression in plate readers. *BMC Biotechnol.* 14:11.

[9] Zulkower, V., M. Page, D. Ropers, J. Geiselmann, and H. de Jong, 2015. Robust reconstruction of gene expression profiles from reporter gene data using linear inversion. *Bioinformatics* 31:i71–i79.

[10] Kannan, S., T. Sams, J. Maury, and C. T. Workman, 2018. Reconstructing dynamic promoter activity profiles from reporter gene data. *ACS Synth Biol* 7:832–841.

[11] Balleza, E., J. M. Kim, and P. Cluzel, 2018. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat Methods* 15:47–51.

[12] Remington, S. J., 2006. Fluorescent proteins: maturation, photochemistry and photophysics. *Curr Opin Struct Biol* 16:714–721.

[13] Verkhusha, V. V., D. M. Chudakov, N. G. Gurskaya, S. Lukyanov, and K. A. Lukyanov, 2004. Common pathway for the red chromophore formation in fluorescent proteins and chromoproteins. *Chem Biol* 11:845–854.

[14] Strack, R. L., D. E. Strongin, L. Mets, B. S. Glick, and R. J. Keenan, 2010. Chromophore formation in DsRed occurs by a branched pathway. *J Am Chem Soc* 132:8496–8505.

[15] Bertero, M., 1989. Linear inverse and ill-posed problems. *Adv Electron Electron Phys* 75:1 – 120.

[16] De Nicolao, G., G. Sparacino, and C. Cobelli, 1997. Nonparametric input estimation in physiological systems: Problems, methods, and case studies. *Automatica* 33:851–870.

[17] Kailath, T., A. H. Sayed, and B. Hassibi, 2000. Linear Estimation. Prentice Hall, Upper Saddle River, New Jersey, USA.

[18] Jazwinski, A. H., 1970. Stochastic Processes and Filtering Theory. Courier Corporation, Massachusetts.

[19] Rasmussen, C. E., and C. K. I. Williams, 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, Massachusetts, USA.

[20] De Nicolao, G., and G. Ferrari-Trecate, 2003. Regularization networks for inverse problems: A state-space approach. *Automatica* 39:669–676.

[21] Davis, J. H., A. J. Rubin, and R. T. Sauer, 2011. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res* 39:1131–1141.

[22] Taniguchi, Y., P. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533–539.

[23] Young, J. W., J. C. W. Locke, A. Altinok, N. Rosenfeld, T. Bacarian, P. S. Swain, E. Mjolsness, and M. B. Elowitz, 2012. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nat Protoc* 7:80–88.

[24] Heux, S., B. Philippe, and J.-C. Portais, 2011. High-throughput workflow for monitoring and mining bioprocess data and its application to inferring the physiological response of *Escherichia coli* to perturbations. *Appl Environ Microbiol* 77:7040–7049.

[25] Shu, X., N. C. Shaner, C. A. Yarbrough, R. Y. Tsien, and S. J. Remington, 2006. Novel chromophores and buried charges control color in mFruits. *Biochemistry* 45:9639–9647.

[26] Miyawaki, A., D. M. Shcherbakova, and V. V. Verkhusha, 2012. Red fluorescent proteins: chromophore formation and cellular applications. *Curr Opin Struct Biol* 22:679–688.

[27] Gibson, D. G., L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison, and H. O. Smith, 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* 6:343–345.

[28] Berthoumieux, S., H. de Jong, G. Baptist, C. Pinel, C. Ranquet, D. Ropers, and J. Geiselmann, 2013. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol Syst Biol* 9:634.

[29] Lambert, T. J., 2019. FPbase: a community-editable fluorescent protein database. *Nat Methods* 16:277–278.

[30] Neidhardt, F. C., P. L. Bloch, and D. F. Smith, 1974. Culture medium for enterobacteria. *J Bacteriol* 119:736–747.

[31] Martin, Y., M. Page, C. Blanchet, and H. de Jong, 2019. WellInverter: a web application for the analysis of fluorescent reporter gene data. *BMC Bioinform* 20:309.

[32] Cinquemani, E., V. Laroute, M. Cocaign-Bousquet, H. de Jong, and D. Ropers, 2017. Estimation of time-varying growth, uptake and excretion rates from dynamic metabolomics data. *Bioinformatics* 33:i301–i310.

[33] Sakamoto, Y., M. Ishiguro, and G. Kitagawa, 1986. Akaike Information Criterion Statistics. Springer Netherlands, Dordrecht.

[34] Stefan, D., C. Pinel, S. Pinhal, E. Cinquemani, J. Geiselmann, and H. de Jong, 2015. Inference of quantitative models of bacterial promoters from time-series reporter gene data. *PLoS Comput Biol* 11:e1004028.

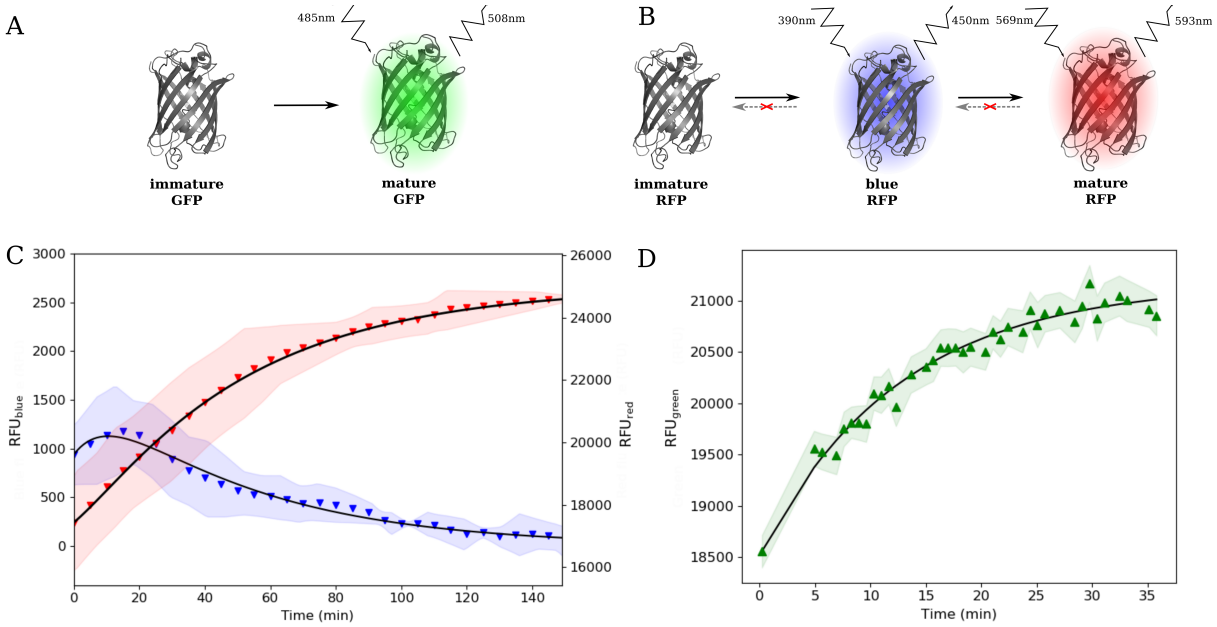

Figure 1: Mechanistic maturation models for RFP and GFP. (**A**) Maturation mechanism of GFP. (**B**) Maturation mechanism of RFP. The possible backflow reactions were eliminated from the model (red crosses). (**C**) Calibration of a mechanistic model of RFP with experimental data: a strain expressing mScarlet-I was grown in MOPS medium supplemented with glucose. At time zero Chloramphenicol was added to the medium to stop translation. Blue and red fluorescence (blue and red triangles respectively) were measured. The plot shows the mean of 6 replicates. Confidence intervals are given by two times the standard error of the mean. These data were used to fit the model of Eqs 3-5 and estimate its parameters (best fit: black solid lines). The plot shows that the mechanistic maturation model captures the maturation dynamics well ($R^2 = 0.91$ for blue fluorescence, $R^2 = 0.99$ for red fluorescence).(**D**) Idem for the calibration of the GFP model of Eqs 1-2 ($R^2 = 0.97$). The absorbance curves are shown in Fig. S1.

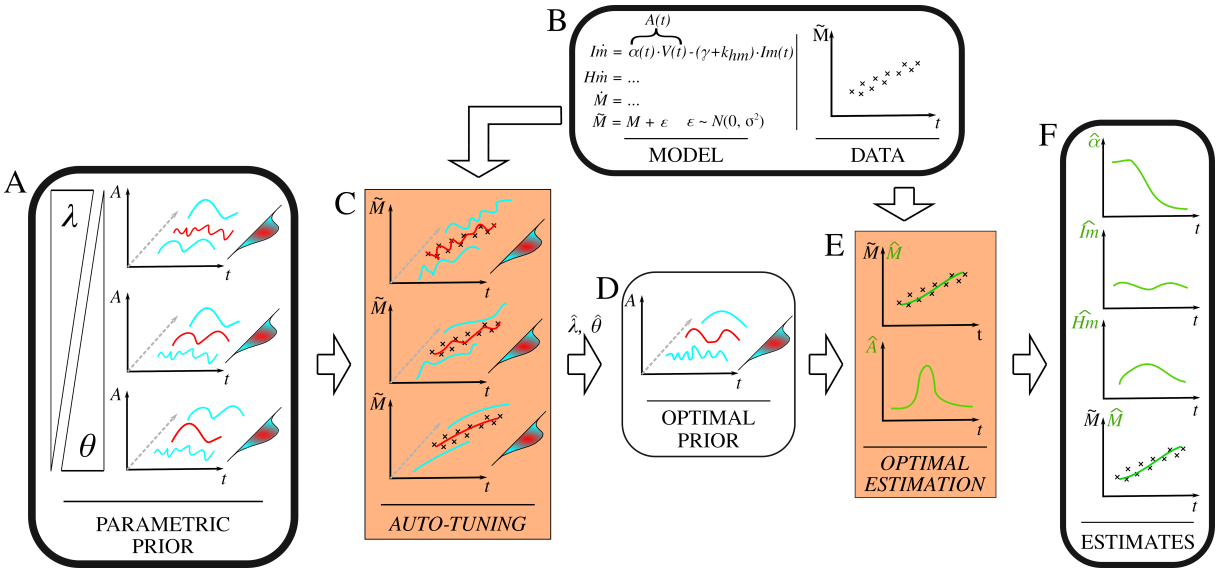

Figure 2: Schematic outline of the Bayesian approach for estimating promoter activities. A parametric family of priors (**A**) expresses expected properties of the profile $A(t)$. Larger values of θ (smaller values of λ) assign higher probability to fast-fluctuating (smaller magnitude) profiles (red solid lines). Together with the gene expression model and fluorescence data (**B**), the auto-tuning step (**C**) selects the best values of θ and λ by evaluating the overall match of the distribution of model-predicted fluorescence profiles with the available data via a maximum-likelihood approach. The resulting optimal prior (*i.e.*, the prior with values for its parameters θ and λ learned from the data) (**D**) is used to produce estimates of the gene expression dynamics and of $A(t)$ via Kalman filtering/smoothing (**E**). Normalization by $V(t)$ eventually yields the estimates of the promoter activity $\alpha(t)$ (**F**). Rounded boxes represent inputs, intermediate results and outputs of the method; rectangular boxes represent procedural steps. The procedure provides robust estimates of promoter activities based on a rigorous, automated selection of regularization parameters θ and λ.

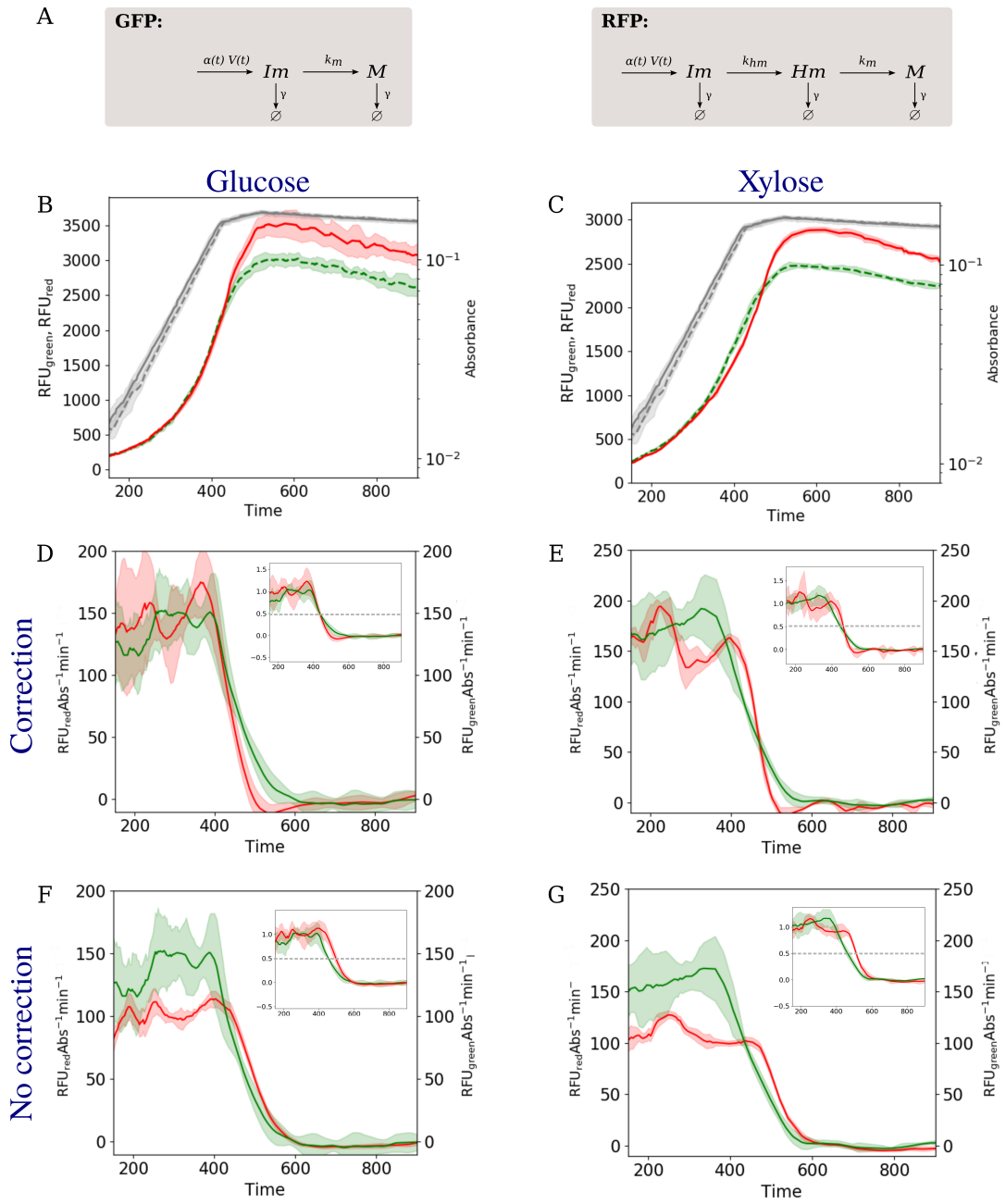

Figure 3: Reconstruction of promoter activities from dynamic fluorescence measurements. (**A**) Model reaction schemes, corresponding to Eqs 1-2 and 3-5, used to reconstruct promoter activities $\alpha(t)$ for GFP and RFP. (**B-C**) Absorbance (grey curves), and red and green fluorescence measurements for a strain expressing either a GFP or an RFP (dotted and solid curves, respectively). Bacteria were grown in a microplate in MOPS minimal medium supplemented with either glucose (B) or xylose (C). The plots show the mean of eight replicates and a confidence interval given by twice the standard error of the mean. (**D-E**) Reconstructed promoter activities for data in glucose and xylose, respectively, for GFP (green) and RFP (red). The inserts show the normalized promoter activities and the value at which they assume the value 0.5 (dashed line). The plots demonstrate that the Bayesian estimation approach yields coinciding promoter activity profiles for green and red FPs. (**F-G**) Idem, but without correction for maturation (Eq. 7). The promoter activities are no longer comparable: the RFP promoter activity is delayed and consistently underestimated in comparison with panels D and E.