BIOINFORMATICS ORIGINAL PAPER
Vol. 31 no. 3 2015, pages 355–362
doi:10.1093/bioinformatics/btu619

Systems biology

Advance Access publication September 30, 2014

Inference of protein kinetics by stochastic modeling and simulation of fluorescence recovery after photobleaching experiments

Maria Anna Rapsomaniki[1,2], Eugenio Cinquemani[3], Nickolaos Nikiforos Giakoumakis[1], Panagiotis Kotsantis[1], John Lygeros[2,*] and Zoi Lygerou[1,*]

[1]Department of Biology, School of Medicine, University of Patras, 26505, Rio, Patras, Greece, [2]Institut für Automatik, ETH Zürich, 8092 Zürich, Switzerland and [3]INRIA Grenoble-Rhône-Alpes, Montbonnot, 38334 Saint-Ismier Cedex, France

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Fluorescence recovery after photobleaching (FRAP) is a functional live cell imaging technique that permits the exploration of protein dynamics in living cells. To extract kinetic parameters from FRAP data, a number of analytical models have been developed. Simplifications are inherent in these models, which may lead to inexhaustive or inaccurate exploitation of the experimental data. An appealing alternative is offered by the simulation of biological processes in realistic environments at a particle level. However, inference of kinetic parameters using simulation-based models is still limited.

Results: We introduce and demonstrate a new method for the inference of kinetic parameter values from FRAP data. A small number of *in silico* FRAP experiments is used to construct a mapping from FRAP recovery curves to the parameters of the underlying protein kinetics. Parameter estimates from experimental data can then be computed by applying the mapping to the observed recovery curves. A bootstrap process is used to investigate identifiability of the physical parameters and determine confidence regions for their estimates. Our method circumvents the computational burden of seeking the best-fitting parameters via iterative simulation. After validation on synthetic data, the method is applied to the analysis of the nuclear proteins Cdt1, PCNA and GFPnls. Parameter estimation results from several experimental samples are in accordance with previous findings, but also allow us to discuss identifiability issues as well as cell-to-cell variability of the protein kinetics.

Implementation: All methods were implemented in MATLAB R2011b. Monte Carlo simulations were run on the HPC cluster Brutus of ETH Zurich.

Contact: lygeros@control.ee.ethz.ch or lygerou@med.upatras.gr

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

Received on October 29, 2013; revised on August 1, 2014; accepted on September 12, 2014

1 INTRODUCTION

The development of modern microscopy systems coupled with advances in fluorescent protein technology and sophisticated computational methods have made it possible to visualize, track and quantify fluorescent molecules within living cells. Functional live cell imaging techniques, such as fluorescence recovery after photobleaching (FRAP), are increasingly used by biology laboratories to explore the dynamic behavior of proteins *in vivo* (Reits and Neefjes, 2001). During a typical FRAP experiment, molecules tagged with a fluorescent protein (such as the Green Fluorescent Protein—GFP) in a subcellular region are irreversibly bleached by a short laser pulse. Then, the recovery of the fluorescence due to unbleached molecules moving into the bleached region is measured by standard time-lapse microscopy. Analysis of this recovery data aims at providing information regarding the diffusion and binding of the bleached molecules, reflecting biomolecular interactions within the cell (Phair and Misteli, 2001).

Conventional quantitative FRAP analysis focuses on parameters associated with the shape of the recovery curve (e.g. value of plateau and half-maximal recovery time), easily estimated using curve-fitting techniques (Phair and Misteli, 2001; Rapsomaniki *et al.*, 2012). However, this approach provides a limited understanding of protein kinetics and is heavily dependent on the experimental setup, especially on the time frame of observation (Bancaud *et al.*, 2010). To estimate the kinetic parameters of the underlying molecular processes, including association and dissociation constants, relative size of mobile and immobile pools and protein diffusion rates, model-based quantitative FRAP analysis is necessary (Phair *et al.*, 2004).

Modeling approaches traditionally rely on developing streamlined models of the diffusion, binding and photobleaching processes to derive approximate closed-form expressions of fluorescence recovery. The parameters yielding recovery curves that correspond best to the data are chosen as the most likely explanation of the underlying protein kinetics (Mueller *et al.*, 2010). Over the past years, a variety of models have been proposed, broadly classified into diffusion models, reaction models and reaction-diffusion models, depending on the phenomenon considered dominant (Sprague and McNally, 2005). It has also been observed (Tardy *et al.*, 1995) that the FRAP recovery curve is composed of two phases: a first phase with fast dynamics, where the recovery is mainly attributed to diffusion (diffusion regime), and a second phase with slower dynamics, where the

*To whom correspondence should be addressed.

recovery is regulated by binding and unbinding events (turnover regime).

Models giving rise to a sum of exponential terms have been extensively used for quantitative analysis. In the reaction models domain, exponential expressions are derived through compartmental modeling approaches pioneered in (Jacquez, 1996), where the presence of one or two exponential terms is dictated by the number of binding sites. An alternative compartmental modeling approach is proposed in (Carrero *et al.*, 2003), which again results in a theoretical curve with two exponential terms. Here, the coefficients of the terms are non-linear functions of all physical parameters (diffusion coefficient, binding/unbinding rates) of the underlying molecular kinetics.

The derivation of explicit (parametric) expressions for FRAP curves relies on simplifying assumptions about cellular/nuclear geometry [2D (Carrero *et al.*, 2003), 3D (Beaudouin *et al.*, 2006)], nature of diffusion [isotropic (Ellenberg *et al.*, 1997) or anisotropic (Sbalzarini *et al.*, 2006)], binding site number [one (Beaudouin *et al.*, 2006) or many (Sprague *et al.*, 2004)] and distribution [homogeneous (Sprague *et al.*, 2004) or heterogeneous (Beaudouin *et al.*, 2006)], to name a few. The accuracy of these approximations is, however, difficult to determine. Furthermore, it has been shown that different models or different parameter sets of the same model can fit FRAP curves equally well. Indeed, contrasting estimates of the kinetic parameters of even the same molecule species have been reported in various studies (Mueller *et al.*, 2010). At the same time, the use of averages of several cell profiles is common practice in traditional FRAP analysis, which may mask the underlying biological information contained in single-cell measurements.

In the past years, the availability of computational resources of ever-increasing power has stimulated modeling and simulation of molecular mobility and interactions at a particle level and within realistic environments (Cowan *et al.*, 2009; Farla *et al.*, 2004; Houtsmuller *et al.*, 1999; van Royen *et al.*, 2009). Stochastic hybrid models, coupling continuous diffusion dynamics with discrete (random) interaction events and providing a realistic account of the complexity of the cellular environment, can be built and simulated in reasonable time (Cinquemani *et al.*, 2008). Analysis of the fit between simulated and experimental recovery curves allows one to (in)validate hypotheses on the values of kinetic parameters. Unfortunately, in this context, a parameter inference method cannot be obtained easily by numerical optimization of the fit, as the repeated simulation of the system for iteratively refined values of the parameters is computationally demanding. To account for this, existing simulation-based methods propose the *a priori* creation of a large dataset of simulated FRAP curves for varying combinations of kinetic parameter values, obtained by gridding the parameter space (van Royen *et al.*, 2009). Parameter inference from experimental data is then performed by searching in the dataset the simulated curve that fits best the experimental one, thus forcing the estimates to take values on the grid of simulated parameters.

In this article, we propose an alternative approach for the inference of the kinetic properties of proteins within living cells. The key idea and novelty of our method is the construction of a mapping from recovery curves to the parameters of the underlying molecule kinetics. The method is based on the numerical simulation of a stochastic hybrid model of protein diffusion and binding in a realistic setup at a particle level (Cinquemani *et al.*, 2008). It assumes that FRAP recovery curves are represented through a sum-of-exponentials. Then, a relatively small number of simulations for different physical parameter values is used to train an Artificial Neural Network (ANN) implementing the required mapping of physical parameters to the parameters of FRAP recovery curves. Given an experimentally determined recovery curve, inference of the underlying kinetic parameters then simply amounts to fitting the data with a sum-of-exponential curve and feeding the obtained recovery parameters to the mapping. In addition, a bootstrap process is used to compute confidence intervals on the parameter estimates. Through the *a priori* construction of the mapping, our method circumvents the burden of resimulating the model iteratively every time a parameter estimate is sought. Unlike existing methods, our method provides interpolation within the range of parameters of the simulated curves. This allows us to reduce the number of simulations required and yet provide estimates that are not biased toward values fixed *a priori*. At the same time, the bootstrap process allows us to investigate the identifiability of the physical model parameters. We validate the proposed method first *in silico* and then through the analysis of *in vivo* data for the nuclear proteins Cdt1, PCNA and GFPnls. Our method predicts behavior in accordance with earlier findings (Essers *et al.*, 2005; Mortusewicz and Leonhardt, 2007; Roukos *et al.*, 2011; Xouri *et al.*, 2007) and provides elements for the discussion of cell-to-cell variability.

2 METHODS

2.1 Model description

We briefly outline the stochastic hybrid model of FRAP experiments, thoroughly presented in (Cinquemani *et al.*, 2008). An outline of the stochastic model is shown in Figure 1, where black boxes represent the

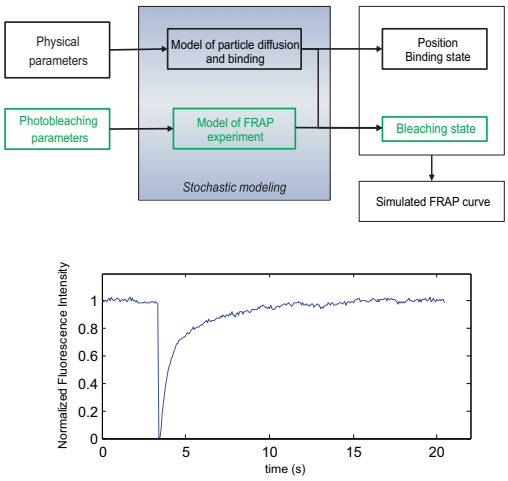

Fig. 1. Above: Outline of the stochastic hybrid model of FRAP experiments. Below: An example simulated curve for a diffusion coefficient of 2 $\mu m^2/s$, a bound fraction of 25% and a residence time of 3 s. A 20 s experiment is simulated with time step 0.068 s leading to 302 FRAP measurements; the approximation parameter (h) used was 0.1 μm

model of particle diffusion and binding and green boxes the model of the photobleaching process.

Protein diffusion and binding are modeled at a particle level, by taking into account explicitly the stochastic nature of diffusion and binding events. For the purposes of this article we have tailored the model to the case of nuclear proteins binding to chromatin and assume that there are no interactions between different particles, i.e. all molecules diffuse and bind independently. In addition, we consider for simplicity that diffusion is isotropic and space-homogeneous and binding and unbinding propensities are uniform over the nucleus. Generalizations of the model are also possible, so that its use is extended for other cases.

We describe the cell nucleus as a 3D ellipsoid with impermeable boundaries, containing N copies of a protein. Let $p_i(t) = [x_i(t)\ y_i(t)\ z_i(t)]^T$, with $i = 1, \ldots, N$ denote the position of molecule i at time t. At each time, every particle is either bound or unbound. Bound molecules do not move, while diffusion of unbound molecules is described by a Brownian motion process through the following stochastic differential equation:

$$dp_i(t) = \sigma I dW_i(t) + dR_i(p_i(t)) \tag{1}$$

where $W(t)$ is a 3D Wiener process with zero mean and covariance equal to the identity matrix I, $\sigma > 0$ and $R(p(t))$ is a process that reflects the molecules back into the nucleus when they would cross the boundaries. In the context of biological systems, the diffusion coefficient of a particle, denoted as D with units $\mu m^2/s$, is associated with a particle's mean square displacement over time. It can be shown (see Section S1.1 of the Supplementary material) that the diffusion coefficient is related to σ through the following equation:

$$D = \sigma^2/2. \tag{2}$$

Transitions of a molecule between bound and unbound states are modeled as random events with propensities $\lambda_{bind} \geq 0$ for binding and $\lambda_{release} \geq 0$ for unbinding (These propensities are equivalent to the association and dissociation rates k_{on} and k_{off}, often found in the biochemical reaction literature.). Let $F \in [0, 1]$ denote the average bound fraction, i.e. the expected fraction of the population of molecules that are bound at any given time. Similarly, let T be the residence time (in seconds), i.e. the time a molecule spends on average in the bound state. Then one has

$$\lambda_{release} = 1/T, \tag{3}$$

$$\lambda_{bind}/\lambda_{release} = F/(1 - F). \tag{4}$$

That is, λ_{bind} and $\lambda_{release}$ determine F and T, and vice versa.

We note here that traditional FRAP analysis involves assessing the immobile fraction (defined as the fraction of bleached molecules that remain in the bleached region at the end of the experiment) and the half-maximal recovery time (denoted as $t_{1/2}$ and defined as the time at which fluorescence intensity within the bleached region equals half of the maximal intensity). Although related, the immobile fraction should not be confused with the bound fraction F defined here; while the first is associated with permanent interactions manifested as plateau values <1 at the end of the experiment and depends on the duration of observation, the latter describes transient as well as permanent interactions depending on the value of the residence time and is not affected by experimental setup. Similarly, $t_{1/2}$ should not be confused with the residence time T, as the first also depends on the speed of diffusion while the second does not. Unlike immobile fraction and half-maximal recovery time, our physical parameters D, F and T explicitly characterize the behavior of the protein of interest and do not depend on experimental settings. For more information on this see Section S2.1 of the Supplementary material.

Overall, the above model results in N independent continuous-time switching diffusions, each describing the position and mobility state of one particle over time. A discrete approximation allows for numerical implementation and simulation of the continuous model. The approximation method is based on the idea of gridding both the state-space and time according to a gridding parameter $h > 0$. Standard results in stochastic analysis (Kushner and Dupuis, 1992) show that the approximate process converges in distribution to the original process as $h \to 0$. The value of h heavily influences the time needed for simulation, with smaller values (thus greater resolution) leading to longer simulation times. For practical purposes, our numerical investigation (Supplementary material, Section S1.2) suggests that the approximation is sufficiently accurate for $h = 0.1$ μm, as no differences in the statistical properties of the resulting simulations are noticed when h is decreased further.

A model of FRAP experiments over an experimental period $[0, \bar{t}]$ involves the model of protein diffusion and binding described above together with a stochastic description of the bleaching process. To model the labeling of proteins with fluorescent tags, we use an additional discrete state (bleaching state) associated to every particle and we assume that initially (pre-bleach time interval) all particles fluoresce. Bleaching is carried out by a continuous laser pulse over a predefined time interval $[t_*, t^*] \subseteq [0, \bar{t}]$ (bleaching interval) in a predefined 3D region inside the nucleus (bleaching region). Following experimentations with various shapes of the bleaching region (see Section S1.4.1 of the Supplementary material), we currently approximate it as a sphere of a fixed radius, positioned in the center of the nucleus. To model the bleaching process, we assume that all particles that enter the bleaching region during the bleaching interval will get bleached with a probability that is proportional to the time spent in the bleaching region and κ_{bleach}, a constant related to the photobleaching efficiency. The value of κ_{bleach}, associated with the intensity of the laser pulse, was determined empirically, so that the bleaching pattern resembles that of experimental data (more details in Section S1.3 of the Supplementary material).

The fluorescence profile inside the bleaching region, denoted by $y(t)$, is obtained by a count of the fluorescent particles inside the region at each measurement time $t \in \{t_1, \ldots, t_M\} \subseteq [0, \bar{t}]$. For normalization purposes, the total number of fluorescent proteins over the whole nucleus, denoted by $Y(t)$, is also recorded at measurement times. The effects of varying the size of the bleaching region and of positioning the bleaching region near the boundaries were also investigated (see Sections S1.4.2 and S1.4.3 of the Supplementary material). Experimental assessment of the simulation procedure was performed by comparison with real data in (Cinquemani *et al.*, 2008).

2.2 Parameter inference method

Let us associate every recovery curve with a vector of parameters $p = (\alpha, \beta, \gamma, \delta)$ taking values in $P \subseteq \mathbb{R}^4_{\geq 0}$. The value of p for a given curve is determined by fitting the model

$$z_p(t) = 1 - \alpha e^{-\beta t} - \gamma e^{-\delta t} \tag{5}$$

to that curve. Details about the fitting process are given in Section 2.2.2; for the moment, we simply assume that every curve corresponds to a single value of p, and that different curves correspond to different values of p. Let $\Theta \subseteq \mathbb{R}^3_{\geq 0}$ be the set of all possible physical parameter vectors $\theta = (D, F, T)$, where D, F and T denote diffusion coefficient, bound fraction and residence time, respectively. In accordance with our stochastic model, several recovery curves, i.e. several parameter values $p \in P$, correspond to the same θ. The region of P associated with a given θ could be obtained by repeated simulation of the model. Alternatively, given one recovery curve with parameters p_0, a cluster of parameters p [denoted as $C(p_0)$] approximating this region is obtained by bootstrapping, i.e. by resampling a family of curves from the p_0 fitting residuals. This strategy applies both to simulated and experimental recovery curves. In the former case, only one simulation of the stochastic model is required, with great computational savings (one simulation takes one to several minutes, depending on θ and the experimental period of interest).

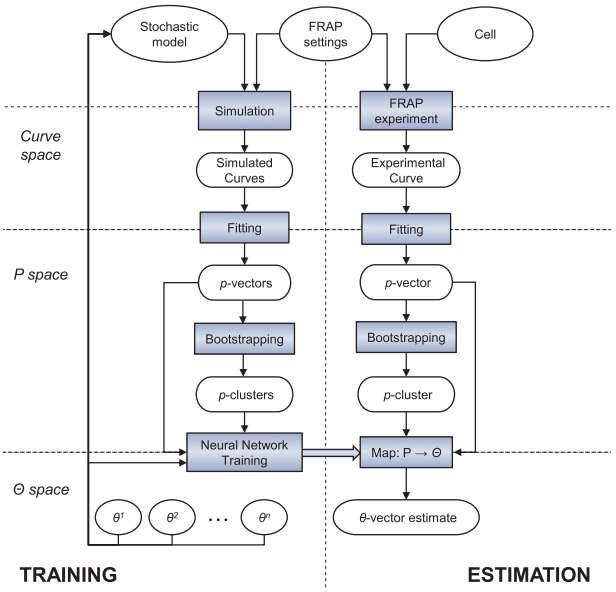

Fig. 2. Outline of the parameter inference method. Left: Training process, executed for many kinetic parameter combinations to generate the desired mapping. Right: Estimation process, executed once per experimental curve to infer kinetic parameters. The methods used for fitting and bootstrapping are common for the training and estimation processes

We can now outline our inference method. The procedure, depicted in Figure 2, is organized in two processes: (i) training process: build a mapping from recovery curves to physical parameters using simulated curves, i.e. from P to Θ and (ii) estimation process: apply the mapping to experimental curves for which an estimate of θ is sought. During the training process (left panel of Fig. 2), we first extract several values of θ randomly distributed over Θ. For each of these values we simulate the model once to get a reference recovery curve, fit the curve with a parametrized function of the form (5) and perform bootstrapping to get a cluster of parameters p. These clusters and the corresponding values of θ are used to train an ANN that maps P into Θ. During the estimation process (right panel of Fig. 2), given an experimental recovery curve, we fit the model of equation (5) to get $\hat{p}$ and perform bootstrapping to get a cluster of parameters $C(\hat{p})$. Then, the mapping constructed during the training process is applied to $\hat{p}$ as well as to every element in the cluster, which yields an estimate of θ (from $\hat{p}$) and a confidence region around it in the form of a cluster of points in Θ (from the cluster of points in P). The accuracy of the method clearly depends on a number of factors and will be discussed based on numerical simulation in Section 3.1. The details of the procedure are given next.

2.2.1 Simulation of FRAP curves To generate recovery curves that represent different FRAP behaviors, we first select randomly n combinations of physical parameter values $\theta^1, \ldots, \theta^n$ in the set

$$\Theta = [0, \overline{D}] \times [0, 1] \times [0, \overline{T}]. \quad (6)$$

Each parameter vector θ^i, $i = 1, \ldots, n$ is then used to simulate a single FRAP curve $y^i(t)$ along with the total nuclear fluorescence intensity $Y^i(t)$ from time 0 to time $\bar{t}$. The resulting sample recovery profiles are normalized using the double normalization process described in (Phair *et al.*, 2004), also used for experimental recovery curves. Normalized recovery curves $z^i(t)$ are defined as follows: for all measurement times $t \in \{t_1, \ldots, t_M\}$,

$$z^i(t) = [y^i(t)/y^i_*]/[Y^i(t)/Y^i_*],$$

where y_* and Y_* are, respectively, the time averages of $y(t)$ and $Y(t)$ over the pre-bleach period $[0, t_*]$. Background fluorescence subtraction, usually carried out on real data before normalization, is not needed for simulated data. Division by total fluorescence corrects for loss of fluorescence due to the photobleaching step, as well as for fluctuations in the fluorescence intensity during the time course of the experiment, due e.g. to acquisition bleaching or fluctuations in laser intensity. Division by prebleach intensities corrects for differences across cells or experiments in the starting intensity in the bleach region relative to the overall nuclear intensity. Such differences may be caused by different cell or bleaching geometries. A direct effect of this normalization process is that, as $t \to \infty$, the normalized curve will rise to plateau values of 1, as fluorescence will progressively become again homogeneous over the whole nucleus. For more details on this and the implications in parameter estimation, see Section S1.4 of the Supplement.

2.2.2 Curve fitting Consider the vector of parameters $p = (\alpha, \beta, \gamma, \delta) \in P$ and the parametric curve $z_p(t)$ defined in equation (5). We use the two-term exponential equation because of its ability to fit well both regimes (diffusive and reaction) typically observed in FRAP curves. The constant term of $z_p(t)$ was set to 1 to reflect the fact that, under full recovery ($t \to \infty$), normalized FRAP curves are expected to plateau to 1 (see Section 2.2.1 above). We fit each (simulated or experimental) curve $z(t)$ as follows. Let $r_p(t)$ denote the residuals $r_p(t) = z(t) - z_p(t)$, with $t = t_m$, $m = 1, \ldots, M$. The vector $\hat{p} \in P$ we associate with the curve is defined as the solution of the following optimization problem:

$$\text{minimize} \sum_{m=1}^{M} r_p(t_m)^2 \text{ with respect to } p \in P$$

$$\text{subject to } \beta > \delta \text{ and } 1 - \alpha - \gamma \geq 0.$$

The first constraint disambiguates the role of the two exponential terms and ensures that the first exponential accounts always for the diffusive (fast recovery) regime. For biological consistency and numerical stability, the second constraint ensures that the fitted curve never goes below zero over the whole post-bleach period (including the time between t^* and the first measurement after it). To solve the resulting non-linear constrained optimization problem, we use the global optimization algorithm OQNLP (Ugray *et al.*, 2006), implemented in MATLAB as the GlobalSearch function.

2.2.3 Bootstrapping Next, we obtain uncertainty clusters using a stochastic approach based on bootstrapping (Efron *et al.*, 1986). For each simulated (or experimental) curve $z(t)$, consider the fitted curve $z_{\hat{p}}(t)$ and the fitting residuals $r_{\hat{p}}(t)$, which we assume to be identically distributed. A new artificial FRAP curve is obtained by bootstrap sampling (sampling with replacements) of the fitting residuals $r_{\hat{p}}(t_1), \ldots, r_{\hat{p}}(t_M)$ and adding them to the fitted curve $z_{\hat{p}}(t_1), \ldots, z_{\hat{p}}(t_M)$. This step is repeated l times and results in l artificial curves, which can be seen as local perturbations of the initial simulated (or experimental) curve, subject to the same amount of noise. Then, by fitting the resulting l artificial FRAP curves as in Section 2.2.2, we construct a cluster $C(\hat{p})$ of l parameters p. This cluster serves as a local estimate of the sensitivity of the p vectors due to noise and process randomness.

2.2.4 Neural network training We construct a simple function-fitting ANN, trained by gradient descent using the Levenberg–Marquardt algorithm, as implemented in the MATLAB Neural Network toolbox. The ANN's architecture is described by an input vector of four features (the entries of the parameter vector p), one hidden layer of 25 neurons and 1 output vector of 3 output elements (the entries of θ). The dataset used for training and testing the ANN comprises the n triplets $(\theta^i, \hat{p}^i, C(\hat{p}^i))$, with $i = 1, \ldots, n$, where each θ^i is one combination of kinetic parameters, sampled as described in Section 2.2.1, $\hat{p}^i$ is the fit of the corresponding

simulated trajectory $z_i(t)$ determined as in Section 2.2.2 and $C(\hat{p}^i)$ is the uncertainty cluster computed from $\hat{p}^i$ as in Section 2.2.3. The input vectors of the ANN are scaled to [–1 1] and the hyperbolic tangent sigmoid is used both as a transfer and activation function. In this way the output range of the ANN is implicitly bounded to [–1 1]. By transforming back to the initial scaling, the output range is bounded to the search space as expressed in (6), preventing the prediction of parameter estimates that lack biological meaning.

The dataset is divided in two subsets: The first (90% of the triplets) is used during the learning phase and the second (the remaining 10% of the triplets) is used as an external test set for *a posteriori* independent assessment of the generalization ability of the trained ANN (ability to predict θ from p for triplets not used in the learning). During the learning phase, the learning dataset is iteratively and randomly partitioned in a training subset (70% of the total triplets) and a validation subset (20%). In each iteration, the training set is used to adjust the network's internal parameters by specifying each θ^i as the output of $\hat{p}^i$ and of all parameters in the corresponding cluster $C(\hat{p}^i)$. The validation set is used to evaluate the training process, which continues until a termination criterion is met (more details in Section S2.3 of the Supplementary material). On execution of this procedure, one gets an ANN that implements the desired mapping $\hat{\theta} : P \to \Theta$.

2.2.5 Implementation We performed a round of 100 Monte Carlo simulations, using as input the physical parameter vectors θ^i, obtained by random sampling in the Θ-space as described in Section 2.2.1. For the purposes of this study, D was sampled in the interval [0, 50] $\mu m^2/s$ by sampling σ uniformly in the interval [0, 10] and using equation (2). T was sampled uniformly in the interval [0, 25] s. To account for the cases of proteins that portray only a diffusive behavior, 10 more θ vectors were added to the dataset with $F = 0$ and D drawn at random in [0, 50] $\mu m^2/s$ (note that T is irrelevant in this case). The total of $n = 110$ *in silico* experiments were simulated; in all cases, the nucleus was represented as an ellipsoid with semi-principal axes of length 5, 4 and 4 μm, the bleaching region as a sphere of radius 2 μm, positioned in the center of the nucleus and the number of particles N was set to 50 000.

Approximation parameter h was set to 0.1 μm for the reasons explained in Section 2.1. Fifty pre-bleach and 250 post-bleach measurements were taken at 0.066 s time intervals and bleaching was attained by a single bleach pulse of 0.066 s. For the above setting, simulation times vary in the order of hours and depend heavily on the choice of the input vector θ^i (for example, fast diffusion demands greater approximation resolution, leading to longer simulation times). The respective simulated curves were fitted as described in Section 2.2.2, yielding 110 p vectors. Following the bootstrap process of Section 2.2.3, $l = 100$ artificial curves were created for each simulated curve and subsequently fitted to obtain the clusters $C(p^i)$.

2.2.6 Parameter estimation from experimental data Parameter inference for an experimental curve is done by repeating the same steps as for the training part, as shown in Figure 2. More specifically, for every given curve y_{obs} we proceed as follows:

(1) Normalize the curve to get z_{obs} (Section 2.2.1);
(2) Fit z_p to z_{obs} and get $\hat{p}_{obs}$ (Section 2.2.2);
(3) Compute the cluster $C(\hat{p}_{obs})$ by bootstrapping (Section 2.2.3);
(4) Get the estimate of θ by feeding $\hat{p}_{obs}$ to the ANN, yielding $\hat{\theta}(\hat{p}_{obs})$;
(5) Get the uncertainty of the estimate $\hat{\theta}$ by feeding the elements of $C(\hat{p}_{obs})$ to the ANN, yielding $\{\hat{\theta}(p) : p \in C(\hat{p}_{obs})\}$.

2.3 Cell culture and FRAP experiments

MCF7 cells were grown in Dulbecco's modified Eagle's medium with 20% fetal bovine serum at 37°C and 5% CO_2 and were transiently transfected with Cdt1-GFP, PCNA-GFP or GFPnls as described in (Roukos *et al*., 2011). FRAP experiments were performed on a Leica SP5 confocal microscope, equipped with a 63×1.4 NA oil immersion lens and FRAP booster. During experiments, cells were plated on Ibidi 30 mm diameter glass-bottom dishes in phenol red-free CO_2-independent medium (Invitrogen) and maintained at 37°C and 5% CO_2. Bleaching of GFP was accomplished on a defined region of interest of 2 μm radius within the cell nucleus. Fifty pre-bleach images were recorded with 4% laser power of the 488 nm line at 40% argon laser intensity, and bleaching was attained by a single bleach pulse of 0.066 s using the 488 and 496 nm laser lines combined at maximum power. After bleaching, 250 images were recorded at 0.066 s time intervals with 4% laser power of the 488 nm line. Raw data were double normalized using the easyFRAP software (Rapsomaniki *et al*., 2012). For FRAP experiments after DNA damage, cells were ultraviolet (UV) irradiated for 10 s (moderate UV dose) using a CL-1000 Ultraviolet Crosslinker UVP and incubated for 1 h before the FRAP experiment.

3 RESULTS AND DISCUSSION

3.1 Performance assessment of the inference method

We first evaluate the performance of the proposed inference method using simulated data. To assess the predictive performance of the ANN in a way that is independent of the specific choice of the test set, we applied a procedure of external k-fold cross-validation. During this process, we repeat the construction of the ANN as described in Section 2.2.4, but for iterative and complementary partitions of the n triplets into learning and testing data. More specifically, the $n = 110$ triplets are randomly partitioned into $k = 10$ subsets (each one of size $n/k = 11$) of which one is set aside as the independent test set and the remaining ones are used as the learning set of the ANN. This process is repeated $k = 10$ times, so that all k subsets are iteratively used as a test set. Note that, for this choice of k, in any iteration the learning and test set represent, respectively, 90 and 10% of the triplets.

In Figure 3 boxplots of the prediction residuals (true minus predicted parameter values) for all $k = 10$, disjoint test sets are plotted. Each boxplot corresponds to all $l = 100$ cluster points of the $n/k = 11$ test samples, i.e. to 1100 prediction residuals. We can see that the ANN's performance for different choices of the training and the test set is comparable, leading to the conclusion that the network training is relatively insensitive to the specific choice of the training data. This fact is crucial for the reliability of our method, which is based on a small simulated dataset. Furthermore, the distribution of the prediction residuals indicates that parameter estimates are effectively unbiased and localized with high confidence near the true parameter values.

3.2 Inference of kinetic parameters from FRAP experimental data

The proposed parameter inference method was used for the analysis of DNA licensing protein Cdt1 and DNA replication/repair protein PCNA. GFP-tagged Cdt1, GFP-tagged PCNA following whole-cell UV irradiation as well as a GFPnls protein were expressed in MCF7 cells and analyzed by FRAP. A total of 14 GFPnls, 16 Cdt1-GFP and 13 PCNA-GFP FRAP curves were analyzed. In Figure 4, the individual FRAP curves (normalized data) for each cell as well as their respective means are shown.

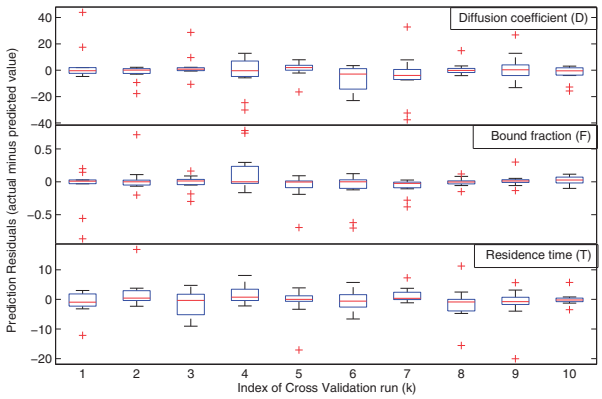

Fig. 3. Boxplots showing the distribution of the prediction residuals (actual minus predicted value) for each parameter and each cross validation run. Diffusion coefficient varies from 0 to 50 $\mu m^2/s$, bound fraction from 0 to 1 and residence time from 0 to 25 s. The red line represents the median, the edges of the box represent the 25th and 75th percentiles and the whiskers extend to the most extreme data points. Outliers are plotted individually as red crosses

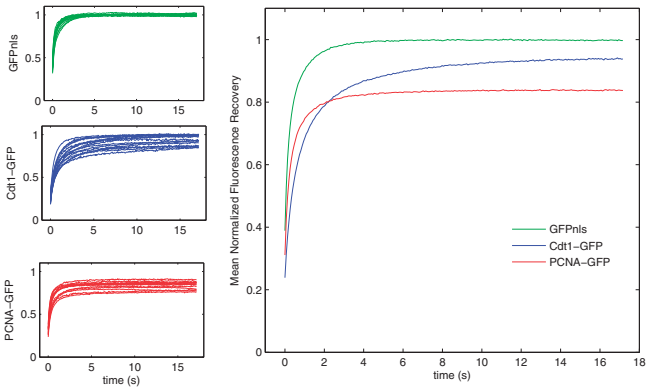

Fig. 4. Experimental FRAP recovery curves (individual curves on the left and their respective means on the right) for GFPnls (green), Cdt1-GFP (blue) and PCNA-GFP (red)

A preliminary analysis of Figure 4 reveals that the different proteins are characterized by markedly different behaviors, both in terms of average behavior (Fig. 4, right) and in terms of variability (Fig. 4, left). Several conclusions can be drawn from the mean curves of Figure 4; the observations that GFPnls exhibits a fast and full recovery, Cdt1-GFP a slow but ongoing recovery and PCNA-GFP a partial recovery hint to their underlying kinetics. The parameter inference method was used to draw estimates of diffusion coefficient, bound fraction and residence time of the three protein species from each curve individually, thus obtaining 14 GFPnls, 16 Cdt1-GFP and 13 PCNA-GFP initial estimates. For each of these, the uncertainty clusters were also computed. The results are shown graphically in Figure 5; numerical values are reported in Supplementary Tables S3–S5.

3.2.1 Predictions agree with known protein behaviors A first observation from Figure 5 is that estimates associated with different proteins localize in different regions of the physical parameter space Θ, separable even by visual inspection. GFPnls estimates are located in the area close to zero bound fraction and zero residence times, indicating a purely diffusive behavior. Estimates for Cdt1-GFP are located in a subspace of Θ where the residence time varies in the order of seconds and the bound fraction varies up to 40%, indicating transient interactions (scanning behavior) coupled to diffusion. Finally, estimates for PCNA-GFP are saturated in the area of higher residence times, suggesting longer immobilization coupled to diffusion for PCNA following UV irradiation. These results agree with qualitative and quantitative analysis of the behavior of these proteins in earlier work (Essers *et al.*, 2005; Mortusewicz and Leonhardt, 2007; Roukos *et al.*, 2011; Xouri *et al.*, 2007).

3.2.2 Estimation of PCNA residence time In agreement with the expectation of a non-negligible fraction of PCNA molecules with longer residence times, predictions for the latter saturate onto the side of Θ corresponding to $\overline{T}$. Motivated by this, we repeated the estimation process for PCNA-GFP by training the ANN with a new simulated dataset, for which the values of T are exponentially distributed in the range [0, 1000] s. As shown in Figure 6, the resulting estimates of T for PCNA-GFP appear to localize in much higher residence times of up to 15 min. At the same time, the expansion of the search space allowed for a better estimation of bound fraction (values of F vary around 50%, which is supported by the biological expectations) and reconfirmed the estimates of D (numerical values reported in Supplementary Table S6). Estimates for the other proteins remained largely unchanged (results not shown), albeit with slightly worse accuracy, which is expected because the much larger search space is explored by the same number of simulated curves.

Regardless of the choice of $\overline{T}$, a further limitation concerning the accuracy of the estimation of T comes from the fact that large values of T (high residence times) are in general difficult to resolve based on the short time span of the experiment (recoveries were followed for 18 s in this dataset). In these cases, longer FRAP experiments would facilitate a more accurate estimation. Simulation results presented in Section S2.2 of the Supplement show that the sensitivity of the recovery curves to different values of T decreases with T itself, and becomes marginal as T gets larger. This sensitivity depends as well on the other physical parameters; in particular it increases with D and F. The latter can be explained considering that for molecules with a higher bound fraction, unbinding events participate more in the recovery, and thus for increasing values of T there is a more noticeable difference in the curves.

3.2.3 Cell-to-cell variability For all three proteins, cell-to-cell variability of the FRAP curves indeed finds its counterpart in the variability of the kinetic parameter estimates, as captured by the different locations of clusters in the Θ space. We observe that clusters corresponding to different GFPnls curves are relatively concentrated (Fig. 5), whereas in the case of Cdt1-GFP (Fig. 5) and PCNA-GFP (Fig. 6) the clusters are more spatially dispersed, indicating higher cell-to-cell heterogeneity. In particular for Cdt1-GFP, the cluster estimates for bound fraction and residence times seem to vary significantly, roughly with inverse proportionality. Because the cells are not synchronized and Cdt1 is a cell cycle regulator, this relation may be owing to changes in the binding behavior of Cdt1. During the cell cycle, non-specific

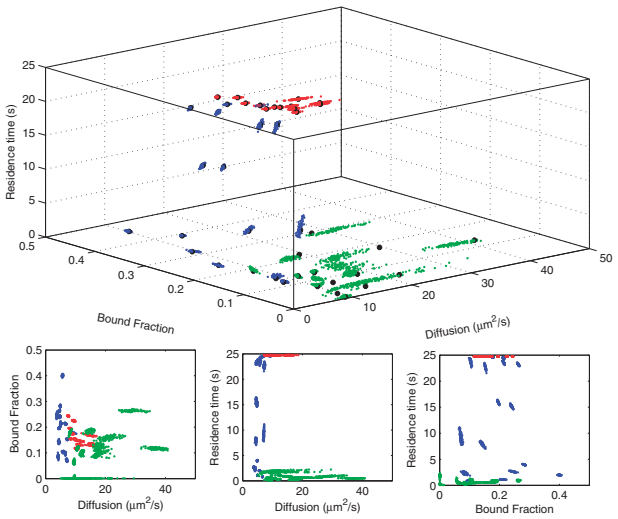

Fig. 5. Cluster estimates for the proteins of interest: GFPnls (green), Cdt1-GFP (blue), PCNA-GFP (red). Top: Parameter estimates shown in the 3D space of kinetic parameters. Estimates for the experimental curves are shown as black dots. Bottom: Respective pairwise projections of the estimates

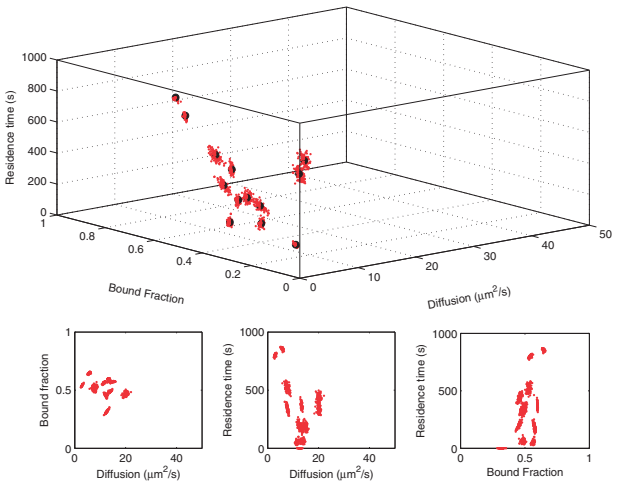

Fig. 6. Predictions for PCNA-GFP for $\overline{T} = 1000$ s reveal residence times up to 15 min. Top: Parameter estimates shown in the 3D space of kinetic parameters. Estimates for the experimental curves are shown as black dots. Bottom: Respective pairwise projections

scanning for binding sites of a larger percentage of Cdt1 molecules could be followed by a more permanent binding of less molecules.

We also notice that variability of FRAP curves may be deceptive: visual inspection of Cdt1 and PCNA curves (Fig. 4, left) suggests a similar amount of cell-to-cell variability, as individual curves seem to spread apart similarly. However, the estimation process reveals a completely different pattern of variability of the Cdt1 and PCNA kinetics, the latter being essentially confined to variable residence times.

3.2.4 Identifiability Finally, we note that both the spread and shape of the estimation clusters is an indication of the identifiability of the underlying kinetic parameters on a single-cell level. In particular, as observed in Figure 5, Cdt1-GFP clusters are relatively symmetrical and concentrated around the estimates of the experimental curve, suggesting that their parameters are identifiable. On the other hand, most PCNA-GFP clusters (Fig. 6) are fairly concentrated in the directions of diffusion and bound fraction, but spread along the dimension of residence time. This is clearly revealed by the confidence intervals reported in Supplementary Table S6 and points to practical identifiability issues concerning this parameter. Inherent difficulties with identifying PCNA residence times can be explained in the light of the discussion in Section 3.2.2.

These findings also show that the sensitivity of FRAP curves, hence the uncertainty of the resulting estimates, is non-uniform over Θ and depends on the actual underlying kinetic parameters. Furthermore, through the construction of the bootstrap clusters, our method allows us to expose identifiability issues, such as when curves are explained by multiple sets of physical parameters. Less clear, but perhaps more intriguing, is whether and how this uncertainty links with the different features of cell-to-cell variability discussed above for the different proteins. This fact is only partially uncovered in the present analysis and deserves future investigation.

4 CONCLUDING REMARKS

In this work, we have proposed a novel kinetic parameter inference method for the quantitative analysis of FRAP data. Our method constructs off-line a mapping between parameters of the FRAP curves and the parameters of an underlying kinetic model. Once this mapping is available, it can be used for inexpensive parameter inference on many FRAP datasets. Thanks to normalization, the same mapping can, to some extent, be applied to varying experimental setups.

Quantitative validation using simulated experiments showed that the method is capable of reconstructing kinetics that were not directly explored via simulation with a reasonable accuracy at an affordable computational cost. When applied to experimental FRAP data for different proteins and from different cells, estimation results were found to be in agreement with existing knowledge. Analysis of the results from individual cells also showed the potential of the method in the study of cell-to-cell variability of protein kinetics and emphasized the importance of single-cell rather than mean recovery curves in this type of analysis.

Through this work, several open problems in FRAP quantitative analysis have emerged. Identifiability and sensitivity of kinetic parameters is a critical issue, which we partially address here by a local analysis based on bootstrapping. On the one hand, a straightforward generalization of the method, based on the use of repeated simulations of the same kinetic parameters at a learning stage, may improve the information on local sensitivity of FRAP curves and resulting estimation uncertainty. However, more work needs to be done so that global information on parameter identifiability is available. Whatever the modeling and estimation method used, quantitative validation of modeling and estimation on experimental data is of critical importance. A big step toward this end would be the availability of benchmark datasets of real experimental data along with accurate and accessible information on the underlying parameter values.

ACKNOWLEDGEMENTS

We thank the Advanced Light Microscopy Facility of the University of Patras.

Funding: This work was supported by a grant from the European Research Council to Z.L. (DYNACOM, ERC StG 281851) and by the European Commission under the HYCON2 Network of Excellence. M.A.R. was supported by a Swiss Government Scholarship.

Conflict of interest: none declared.

REFERENCES

Bancaud,A. *et al.* (2010) Fluorescence perturbation techniques to study mobility and molecular dynamics of proteins in live cells: FRAP, photoactivation, photoconversion and FLIP. *Cold Spring Harb. Protoc.*, **2010**, pdb.top90.

Beaudouin,J. *et al.* (2006) Dissecting the Contribution of Diffusion and Interactions to the mobility of nuclear proteins. *Biophys. J.*, **90**, 1878–1894.

Carrero,G. *et al.* (2003) Using FRAP and mathematical modeling to determine the in vivo kinetics of nuclear proteins. *Methods*, **29**, 14–28.

Cinquemani,E. *et al.* (2008) Numerical analysis of FRAP experiments for DNA replication and repair. In: *Proceedings of the IEEE Conference on Decision and Control*. Cancun, Mexico, pp. 5180–5185.

Cowan,A.E. *et al.* (2009) Using the virtual cell simulation environment for extracting quantitative parameters from live cell fluorescence imaging data. *Microsc. Microanal.*, **15**, 1522–1523.

Efron,B. and Tibshirani,R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–75.

Ellenberg,J. *et al.* (1997) Nuclear membrane dynamics and reassembly in living cells: targeting of an inner nuclear membrane protein in interphase and mitosis. *J. Cell Biol.*, **138**, 1193–1206.

Essers,J. *et al.* (2005) Nuclear dynamics of PCNA in DNA replication and repair. *Mol. Cell. Biol.*, **25**, 93509359.

Farla,P. *et al.* (2004) The androgen receptor ligand-binding domain stabilizes DNA binding in living cells. *J. Struct. Biol.*, **147**, 5061.

Houtsmuller,A.B. *et al.* (1999) Action of DNA repair endonuclease ERCC1/XPF in living cells. *Science*, **284**, 958961.

Jacquez,J.A. (1996) *Compartmental Analysis in Biology and Medicine*. 3rd edn. Biomedware, Ann Arbor, MI, p. 512.

Kushner,H.J. and Dupuis,P.G. (1992) *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer, New York, NY.

Mortusewicz,O. and Leonhardt,H. (2007) XRCC1 and PCNA are loading platforms with distinct kinetic properties and different capacities to respond to multiple DNA lesions. *BMC Mol. Biol.*, **8**, 81.

Mueller,F. *et al.* (2010) FRAP and kinetic modeling in the analysis of nuclear protein dynamics: what do we really know? *Curr. Opin. Cell. Biol.*, **22**, 403–411.

Phair,R.D. and Misteli,T. (2001) Kinetic modelling approaches to *in vivo* imaging. *Nat. Rev. Mol. Cell Bio.*, **2**, 898–907.

Phair,R.D. *et al.* (2004) Measurement of dynamic protein binding to chromatin *in vivo*, using photobleaching microscopy. *Method Enzymol.*, **375**, 393–414.

Rapsomaniki,M.A. *et al.* (2012) easyFRAP: an interactive, easy-to-use tool for qualitative and quantitative analysis of FRAP data. *Bioinformatics*, **28**, 1800–1801.

Reits,E.A.J. and Neefjes,J.J. (2001) From fixed to FRAP: measuring protein mobility and activity in living cells. *Nat. Cell Biol.*, **3**, 145–145.

Roukos,V. *et al.* (2011) Dynamic recruitment of licensing factor Cdt1 to sites of DNA damage. *J. Cell Sci.*, **124**, 422–434.

van Royen,M.E. *et al.* (2009) Fluorescence recovery after photobleaching (FRAP) to study nuclear protein dynamics in living cells. *Methods Mol. Biol.*, **464**, 363–385.

Sbalzarini,I.F. *et al.* (2006) Simulations of (An)Isotropic diffusion on curved biological surfaces. *Biophys. J.*, **90**, 878–885.

Sprague,B.L. *et al.* (2004) Analysis of binding reactions by fluorescence recovery after photobleaching. *Biophys. J.*, **86**, 3473–3495.

Sprague,B.L. and McNally,J.G. (2005) FRAP analysis of binding: proper and fitting. *Trends Cell. Biol.*, **15**, 84–91.

Tardy,Y. *et al.* (1995) Interpreting photoactivated fluorescence microscopy measurements of steady-state actin dynamics. *Biophys. J.*, **69**, 1674–1682.

Ugray,Z. *et al.* (2006) Scatter search and local NLP solvers: a multistart framework for global optimization. *INFORMS J. Computing.*, **19**, 328–340.

Xouri,G. *et al.* (2007) Cdt1 associates dynamically with chromatin throughout G1 and recruits geminin onto chromatin. *EMBO J.*, **26**, 1303–1314.