Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 23 Oct 2016
Manuscript Category

OXFORD

Computational and Systems Biology

Model inference from protein time-course in Hematopoietic Stem Cells (HSC)

Pandu Raharja [1,2,3,*], Rene Schoeffel [1,2,3], Michael Strasser [3] and Carsten Marr [3,*]

[1] Technische Universität München, Fakultät für Informatik, Boltzmannstr. 3, 85748 Garching bei München, Germany
[2] Ludwig-Maximilians-Universität München, Professor-Huber-Platz 2, 80539 München, Germany
[3] Helmholtz Zentrum München, Institute of Computational Biology, Ingolstädter Landstr. 1 85764 Neuherberg, Germany.

* To whom correspondence should be addressed.

Associate Editor: Jan Quell

Received on 23/09/2016; revised on 14/10/2016; accepted on 21/10/2016

Abstract

Motivation: In single cell setting, the dynamics of the systems governing the cell – gene expression being one – are heavily influenced by the stochasticity of each involving element. These stochastic dynamics in gene expression appears to convey more information about the underlying mechanism of gene expression processes than it would be otherwise assumed through the study of population average.
Results: In this paper we presented a particle filtering-based algorithm and a framework that is capable of inferring parameters underlying the stochastic models from single cell expression data. This framework was then applied on time-lapsed microscopy data of two transcription factors (`Pu.1` and `Gata.1`) in murine blood stem cells. It is thought that both transcription factors play decisive roles in several stages of stem cell differentiation, especially the differentiation of myeloid progenitors. Our results provide several insights into the dynamics of blood stem cells maturation and specifically in the single cell environment. We managed to gain several valuable insights on the interaction dynamics between two transcription factors and the outcome of cell maturation. While doing so, we managed to developed highly flexible general purpose highly parallelized framework to model dynamical systems using particle filtering. The general framework software is now available as free software for anyone to use.
Availability: The data are available upon request. A general framework for determining the parameters that best explain the data were published under open source license. The complete source code of the program is publicly accessible at https://github.com/raharjaliu/PFInfer.
Contact: pandu.raharja@tum.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advances in gene expression measurement techniques, coupled with equally rapid advances in single cell analytics, have allowed the more recent high resolution single cell analysis of cell dynamics (Feigelman, 2016; Hoppe *et al*, 2016). This has enabled us, for example, to look deeper into the dynamics of stem cell maturation. In recent years, several genes that potentially are involved in the decision making mechanism of cell maturation going from Inner Mass Cell (IMC) all the way to somatic cells have been discovered and widely known by now (Graf & Enver, 2009). While this has enable us to study cell processes more thoroughly, the deterministic population-averaged nature of cell dynamics was becoming less and less apparent while the stochastic nature of single cell dynamics was becoming dominating as the analysis went down onto single cell resolution (Elowitz *et al*, 2016). Hence our motivation to apply inference methods that are capable to infer dynamics of the systems that are heavily influenced by non-deterministic factors.

For this project we use particle filtering to learn more about aforementioned dynamics. We apply this technique on our time-dependent expression data of single cells containing concentration of both transcription factors. The expression data were measured from time-lapsed microscopy of fluorescent-tagged proteins' expression level of the cells.

The approach of the experiment would be further described in the next section followed by the theoretical aspects of the particle filtering and other methods used in this project.

In this paper we focus on the maturation lines of hematopoietic stem cells (Orkin *et al*, 2008). The cell lines start with Long Term Hematopoietic Stem Cells (LT-HSC), which is the progenitor of all blood cells. During its life, LT-HSC could either divide into two LT-HSC or turn into Short Term Hematopoietic Stem Cell (ST-HSC). SH-HSC will then undergo a transition process into either Common Myeloid Progenitor (CMP) or Common Lymphoid Progenitor (CLP). As the names suggest, both cells are the ancestor of all myeloid cells and lymphocites, respectively. Each would then further maturate into their respective intermediate cells and eventually the mature somatic cells (Red Blood Cells, Megakaryocytes, Mast Cells, Eosinophils, Neutrophils and Macrophages from CMP and B and T Lymphocytes from CLP; see Figure 1).

The decision process of CMP transitioning into either MEP and GMP is the focus of our project. Two transcription factors, `Pu.1`and `Gata.1` are thought to influence the decision process (Graf & Enver, 2009). Moreover, both transcription factors are known to inhibit each other. It is thought that the whole cross inhibition dynamics between the two has an impact on the fate of Common Myeloid Progenitor. This dual-agent dynamics is thought to behave in a way that is self-fulfilling. That is, a slight change in concentration of transcription factors favoring one state over another will have stupendous impact on cell fate decision (Zhang *et al*, 2003).

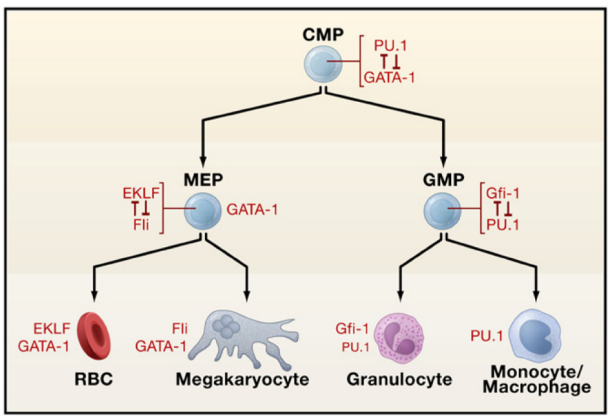


Fig. 1. Maturation cascade of intermediate Common Myeloid Progenitor (CMP) to Megakaryocyte-Eerythroid Progenitor (MEP) and Granulocyte/Macrophage Progenitor (GMP). MEP would then maturate into somatic Red Blood Cells and Megakaryocytes while GMP would undergo maturation into Granulocytes and Monocyptes. For each process there are known factors which are known to interact antagonistically on maturation decision. Our CMP to MEP and GMP decision stands on the top of the figure with two transcription factors of interest, `Pu.1` and `Gata.1` interact atagonistically. Figure taken from Graf & Enver, 2009.

2 Approach

For the single cell dynamic analysis to be possible, we were interested in gaining insight in single cell transcription data of both `Pu.1` and `Gata.1`. We combine two techniques for this to be possible: single cell tracking and protein tagging.

2.1 Experimental Setting

The data for the experiment were taken from Hoppe *et al*, 2016 (Hoppe *et al*, 2016). The expressed protein of interests `Pu.1` and `Gata.1` were tagged with distinct fluorescence proteins `eYFP` and `mCherry` respectively. Time-resolved expression levels were then measured based on the intensity of aforementioned florescence proteins.

Time-lapse imaging was performed at 37 °C in fibronectin coated channel slides provided by Takara Bio. The light source for the experiment was a HXP 120 from Zeiss. 46HE and 43HE, both provided by Zeiss, were used to detect eYFP and mCherry fluorescence coming from the cells at exposure times between 400-1,500 ms using an AxioCam HRm from Zeiss. Bright field pictures were taken every 60-120 s while fluorescent pictures used for the measuring `PU.1-eYFP` and `GATA1-mCherry` were acquired every 30 min. Pictures used for analysis were saved in lossless TIF or PNG format. Single-cell tracking and image quantification were performed using self-written software as described by Hoppe *et al*, 2016.

Figure 2 shows how the dataset from similarly conducted experiment taken from Feigelman, 2016

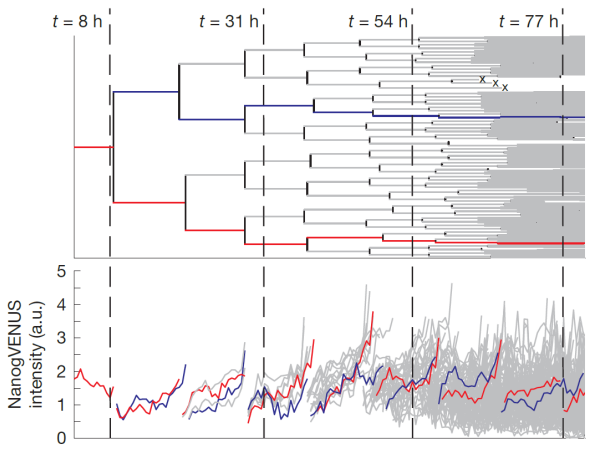


Fig. 2. Example of single cell expression data. In this example, initially only one cell is observed. The cell would then divide into two cells. The system automatically recognizes this and separate measurements of expression would then be conducted, which is visible in the lower part of the figure. The above example also emphasizes the uniqueness and the tracability of each cell by pointing two specific lineages within the data: the blue and red lineages. The system records complete expression measurement of the lineages which could be seen in second half of the figure corresponding with the first half of the figure. The system is capable of recording not only cell division but also cell deaths, as can be seen through the x's sign in the upper part of the figure. Figure taken from Feigelman, 2016.

3 Models

We utilize particle filtering to learn the dynamics of maturation of Common Myeloid Progentior (CMP) cells. The definition of particle filtering could be seen in the subsection **Particle Filtering**. **Simulation Process** details how the simulation is done and the probabilistic approaches is used in it. The assumed model of the cell maturation is then described in the subsection **Reaction Models**. A framework for model comparison is also presented in subsection **Models Comparison**. This is especially helpful for comparing the default model described in **Reaction Models** and other possible alternative models.

3.1 Particle Filtering

Particle filtering, which is also known as **Sequential Monte Carlo** (SMC) (Doucet *et al*, 1945; Liu and Chen, 1998)), is a class of methods generally used to solve the filtering problem. Classically, filtering problem is the

problem of identifying best estimate of some system given noisy and incomplete input – hence the word filter (Del Moral, 1996). As the alternative name implies, it is a set of methods that run Monte Carlo sequentially interrupted by inference phase. It consists mainly of two phase steps, **prediction-updating** and **mutation-selection**, that are done sequentially and repetitively. In the first phase, the prediction-updating updates the conditional probability of underlying assumption – generally referred to as **posterior distribution** – of a system given the previous simulation. Later on, a simulation (mutation) is done on given underlying assumption and simulations best fit the data would then be selected to update the underlying assumption in the next prediction-updating step (Del Moral, 2012).

As the name also implies, particle filtering consists of sequentially and repetitively filtering of mathematical construct called particle, an abstract representation of combination of model, data and parameters. In the next part we will dive deep into the theory of particle filtering. First, we will define what a particle is. This is then followed by definition of Propensity, Posterior, Parameters and Prior. Finally the section will be closed by short review of other methods that also tried to stochastically infer parameters on time-lapsed fluorescent single cell data.

3.1.1 Particle

A particle $\mathcal{K}$ is defined as a triple of trajectory X, parameter set θ and assumed model $\mathcal{M}$,

$$\mathcal{K} := (X, \theta, \mathcal{M}) \tag{1}$$

Specifically in our case, trajectory refers to the concentration of measured proteins `Pu.1` and `GATA1` while parameter set encompasses all variables that were involved in the reaction such as propensity and its corresponding auxiliary variables (see subsection **Reaction Models**). Note that a trajectory is different from the experimental data $\mathcal{D}$. A trajectory is the result of the simulation done by applying the parameters onto our model. In functional notation we would assume the i-th value of the trajectory X_i to be a function of the i-th value of the trajectory, the model and the parameters,

$$X_i = f(X_{i-1}, \mathcal{M}, \theta_{i-1}) \tag{2}$$

3.1.2 Propensity

As the name suggests, propensity is measure of how likely a certain thing would be happening in a given time. In probabilistic theory, it is understood as a tendency a certain type of physical situation would yield an outcome of certain kind (Nodelman *et al*., 2003). In our model specifically, a propensity roughly corresponds to the chance of a reaction to happen in a given time,

$$a_i \propto P(R_k = r_i) \tag{3}$$

Where $P(R_k = r_i)$ refers to the probability of reaction at time point k to be the -th reaction. For reactions $r_1, r_2, \cdots r_p$ there are propensities $a_1, a_2, \cdots a_p$ that are uniquely assigned to each reaction. The probability of a reaction r_i to happen in a given time is thus the ratio between its propensity a_i to the total sum of all propensities (Gillespie, 1977),

$$P(R_k = r_i) = \frac{a_i}{\sum_{j=0}^{p} a_i} \tag{4}$$

In our model, two kinds of propensities were defined. For uninhibited reaction, i.e. a reaction in which no inhibiting effect from other species – a term denoting any entity that is involved in a reaction network – is assumed, the reaction specific propensity is defined as the parameter in the mass action law of chemical reaction:

$$a_i = k_i \tag{5}$$

with k_i referring to the reaction parameter of r_i.

For reaction r_i inhibited by a species Y, Michaelis-Menten inhibition kinetics is assumed to be the propensity of the reaction (Michaelis & Menten, 1913),

$$a_i = k_i \cdot (1 - \frac{X_i^n}{X_i^n + K_Y^n}) \tag{6}$$

where k_i, X_i and K_Y refer to reaction parameter of r_i, concentration of product X in a given time and inhibiting coefficient of regulator Y on X respectively.

3.1.3 Posterior

Our posterior describes the probability of having the trajectory X and parameter θ given the observation $\mathcal{D}$,

$$P(X, \theta|\mathcal{D}) \tag{7}$$

This is understood as the probability of having our simulation return a given set of values *and* having the parameter set θ given that we previously observed the experimental data $\mathcal{D}$. Using Bayes' Theorem we could further expand our posterior into an update rule,

$$P(X, \theta|\mathcal{D}) = \frac{P(\mathcal{D}|X, \theta)P(X, \theta)}{P(\mathcal{D})} \tag{8}$$

In the simulation it is well known that, to compute prior $P(\mathcal{D}|X, \theta)$, only knowledge about the trajectory of the simulation is needed (Feigelman, 2016). Hence, the prior could be simplified as,

$$P(\mathcal{D}|X, \theta) = P(\mathcal{D}|X) \tag{9}$$

Note that the above equation inherently assumes that $\mathcal{D}$ is only directly dependent on X and X is in turn only directly dependent on θ. Incorporating this onto our update rule, and expanding Equation (8) using chain rule, we get

$$P(X, \theta|\mathcal{D}) = \frac{P(\mathcal{D}|X)P(X|\theta)\pi(\theta)}{P(\mathcal{D})} \tag{10}$$

One of the interesting aspects of our method is the fact that the i-th simulation result is only dependent on previous simulation, a property known as *Markov property*. We could thus rewrite the update rule as follow,

$$P(X, \theta|\mathcal{D}) = \frac{\prod_{i=0}^{N} P(\mathcal{D}_i|X_i)P(X_0)\prod_{l=1}^{N} P(X_{[tl-1,tl]}|X_l, \theta)\pi(\theta)}{P(\mathcal{D})} \tag{11}$$

3.1.4 Parameters

During the simulation we assumes certain parameters that would influence the trajectory of the simulation. The parameter set θ is a P-tuple containing all the parameters that are assumed in the simulation,

$$P := (P_0, P_1, \ldots, P_P) \tag{12}$$

3.1.5 Prior

Our prior $\pi(\theta)$ could be expanded by assuming the independence of each parameter θ_i within the parameter set θ,

$$\pi(\theta) = \prod_{i=1}^{P} \pi(\theta_i) \tag{13}$$

Specifically for this simulation, we assume our prior to be gamma distributed with the parameters α and β,

$$\pi(\theta) = \prod_{i=1}^{P} Ga(\theta_i, \alpha_i, \beta_i) \tag{14}$$

We used gamma distribution as our prior in our model due the fact that the, according to (Feigelman, 2016), the posterior of parameters would be in turn be gamma distributed, a fact which will be shown later in results section. It is thus computationally convenient to estimate the posterior if gamma prior is assumed. This phenomenon in which prior and posterior having the same distribution family is known in Bayesian Statistics as *conjugate prior* (Gelman *et al*., 2009).

3.1.6 Comparison with other methods

There are several methods that infer stochastic processes. Wilkinson – and later on, Golightly and Wilkinson – inferred stochastic differential equation (SDE) kinetic model of bacterial gene regulation using likelihood-free Markov Chain Monte Carlo (MCMC) on florescence data (Wilkinson, 2010; Golightly and Wilkinson, 2010). Gonzales *et al* did two stochastic inference methods, Mixed Effects (ME) and the Chemical Master Equation (CME), on time-lapsed single-cell data. The paper also tested the method by inferring previously known HOG pathway in yeast *Saccharomyces cerevisiae* (Gonzales *et al*., 2013). Koromowski *et al* developed Bayesian framework performed using Markov Chain Monte Carlo (MCMC) and linear noise approximation to calibrate models on fluorescent experiment data (Komorowski *et al*., 2009).

3.2 Inference Process

The simulation is run in the following main steps:

1. Initialization of parameters θ.
2. Input of data $\mathcal{D}$.
3. Particle filtering routine:
   a. Generation of initial particles for step i

$$Ki := (K_{i1}, K_{i2}, \ldots, K_{im}) \tag{15}$$

   b. Simulation run of each particle K_{ij}
   c. Weighting of each particle. The weight is a function of the probability of observing the data given the simulation result.

$$w_i^k = P(D_i|X_i^k) = \mathcal{N}(\mathcal{D}_i|X_i^k) \tag{16}$$

   d. Parameter update for every K,

$$\theta^k \propto P(\theta|X^k_{[to,ti]}) \tag{17}$$

4. Model comparison.

The visualization of the whole particle filtering process could be seen in Figure 3.

3.2.1 Particle Generation

As mentioned before, A particle $\mathcal{K}$ is defined as a 3-tuple of trajectory X, parameter set θ and model $\mathcal{M}$. For a specific model, j-th particle at i-th time particle is then just a tuple of trajectory and parameter set $k_{ij} = (x_i, \theta_{ij})$.

For initial time, a particle could be constructed by combining initial parameters, which were arbitrarily defined by a pre-defined prior, and initial trajectory value X_0. There are three ways to define the initial trajectory value. First is to take 0 as initial value. Second is to take the first measurement data $\mathcal{D}_0$. Third is to model initial data as normal distributed instances around $\mathcal{D}_0$, i.e. $X_0 \propto \mathcal{N}(\mathcal{D}_0)$. It is worth noting that the first strategy may not always be applicable.

M particles would then be created and for each iteration of the simulation with each particle having weight which calculated from

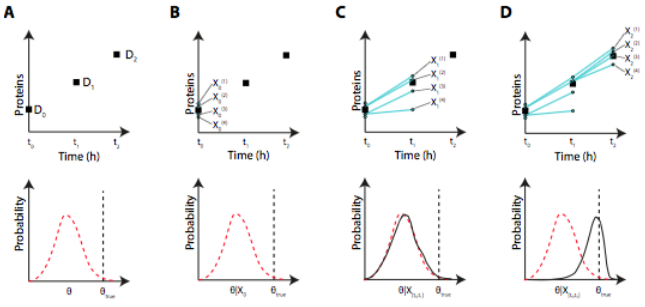


Fig. 3. Visualization of particle fil tering. (A): The particle filter requires a series of observation $\mathcal{D} = (\mathcal{D}_0, \mathcal{D}_1, \cdots \mathcal{D}_N)$ (top) and a prior distribution for model parameters $\pi(\theta)$ (bottom) as input. (B): Particles are initialized by sampling initial states $X(K)$ for each particle K. Parameters Θ are sampled from the prior distribution $\pi(\theta)$. (C): The trajectories are resampled according to the likelihood $w_0^{(k)} = P(D_0|X_0(k))$ and propagated to the next time step using stochastic simulations to generate new states $X_1^{(k)}$ at timepoint t_1. Model parameters for each latent trajectory are resampled from the conditional distribution $P(\theta|X^{(k)}_{[t0,t1]})$. (D): At each iteration, the weights are recomputed and the particles are resampled. Resampled particles are propagated to the next timepoint and the parameters are resampled conditional on the resampled latent trajectories. Over time the posterior parameter distribution converges to the true value. Figure taken from Feigelman, 2016 Feigelman, 2016

previous iteration w_{i-1}^k. In the next iteration, M particles are to be chosen. Note that this implies that a particle from previous iteration could be chosen more than once. Corollarly, a particle from previous iteration may also not be chosen at all. The draw is done by first drawing a random $z \propto \mathcal{U}(0,1)$. A particle k_{ij} is chosen with j denoting the smallest index so that the sum of all weight of particles with index smaller than j is larger than z, i.e.

$$\sum_{l=1}^{j} w_{i-1}^l \geq z \tag{18}$$

3.2.2 Simulation Run

Simulation is done using Gillespie's SSA algorithm (Gillespie, 1977). Dating back to as early as 1945, when it was developed by Joseph Doob Doob, 1945 Chung, 1967, modern Gillespie implementation includes improvements developed years after the invention of the algorithm such as τ-leaping and Bayesian Approximation Method. Such improvements are essential since they would reduce computational cost of the simulation and in turn enable the simulation to scale further to accommodate more complex simulation involving larger systems – in our case we could simulate more cells at the same time in possible more fine-grained time-resolved manner. In our case especially, such improvements are needed to accommodate the explosion of the number of cells within our culture. Specifically, an approximative method would prevent a bottleneck in simulation caused by the exponential increase of the number of living cells that have to be simulated.

Tau Leaping

Generally, tau-leaping works by performing all reactions happening for an interval of length tau before updating the parameters and propensity function (Cao *et al*., 2003). By introducing this kind of approximation we allow more efficient simulation and thus increase the capability of the simulation to cope with larger systems. In our case, we use following tau,

$$\tau = \frac{1}{a_0} ln\Gamma \tag{19}$$

with a_0 referring to total sum of reaction propensities and Γ being uniformly distributed between 0 and 1.

3.2.3 Particle Weighting

Upon the completion of all simulations within one iteration, the quality of each simulation (and in turn, the particle) will be quantified. This quantification is implied in the weighting w_i^k of the particle at the next iteration. We assume the Gaussian distribution of particle around the measurement time at time $t = i$. Using this assumption, we can quantify our weight in following manner,

$$w_i^k = P(\mathcal{D}_i|X_i^k) = \mathcal{N}(\mathcal{D}_i|X_i^k) \tag{20}$$

3.2.4 Parameters Update and Gamma Distribution

After each run, the parameters would then be updated using the underlying Gamma Distribution. For k-th particle, the updated parameters at time i are dependent on previous parameters given the trajectory of data in $[t_0, t_i]$,

$$\Theta^k \propto P(\Theta|X_{[t_0,t_i]}) \tag{21}$$

Applying Bayes' Rule on the probability and independent assumption, we get following equation,

$$P(\Theta|X) = \frac{P(X|\Theta)}{P(X)} \prod_i^P P(\Theta_i) \tag{22}$$

The Gamma Distribution is particularly favored for the update since a conjugate prior of Gamma Distribution is in turn Gamma distributed,

$$P(\Theta|X) = \frac{P(X|\Theta)}{P(X)} \prod_{d=1}^P Ga(\Theta_d, \alpha_d, \beta_d) \tag{23}$$

With α and β referring to both the shape and rate of the distribution respectively. Owing to the fact that conjugate prior of Gamma Distribution is also Gamma distributed, this could be then further simplified as an update rule of the Gamma distribution,

$$P(\Theta|X) = \frac{P(X|\Theta)}{P(X)} \prod_{d=1}^P Ga(\Theta_d, \alpha_d + r_d, \beta_d + Gd) \tag{24}$$

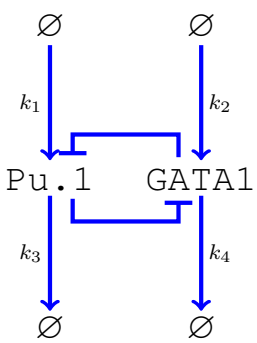
r_d here refers to the number of times the d-th reaction was fired during the run and G_d is defined as follows,

$$G_d := \frac{1}{k_d} \sum_{l=0}^{i} a_l(X(S)) \tag{25}$$

Here, k_d and $a_l(X(S))$ refer to reaction rate of the d-th reaction and trajectory $X(S)$-dependent propensity function, respectively.

3.3 Reaction Models

We assume following putative cross inhibition between genes `Pu.1` and `Gata.1` playing roles in cell differentiation dynamics:



There are four possible reactions happening within our model: k_1, k_2, k_3 and k_4. In our simulation, each reaction was assigned propensity value which roughly is proportional to the likelihood of the reaction happening in a given time:

$$P(R_i = k_l) \propto a_l \tag{26}$$

The propensity a of decay reactions k_3 and k_4 roughly follows the mass action law of chemical reaction:

$$a_l = k_l \cdot A;\ l \in [3, 4] \tag{27}$$

with A_3 and A_4 referring to number of `Pu.1` and `GATA1` respectively. For production reactions k_1 and k_2 we assume Michaelis-Menten inhibition kinetics that influences the production of the species. In this assumption, the inhibiting characteristics of a species on another species negatively influences the creation rate of the species it inhibits. The propensity of the reactions are thus,

$$a_1 = k_1 \cdot (1 - \frac{\texttt{GATA1}^n}{\texttt{GATA1}^n + K_{P1}^n}) \tag{28}$$

$$a_2 = k_2 \cdot (1 - \frac{\texttt{Pu.1}^n}{\texttt{Pu.1}^n + K_{P2}^n}) \tag{29}$$

With K_{P1} and K_{P2} referring to the respective protein concentrations at which inhibiton trough `GATA1` or `Pu.1` is at half its maximum effect.

3.4 Models Comparison

Our framework enables us to test not only standard hematopoietic stem cell maturation as described above, it also allows us to compare it different models. One method that could be used to compare models is Bayes Factors. It is done by performing the ratio of the posterior probabilities of two models M_1 and M_2,

$$B_{M1,M2} = \frac{P(M_1|\mathcal{D})}{P(M_2|\mathcal{D})} \tag{30}$$

Using Bayes' Theorem, we can formulate the marginal probability as,

$$P(M|\mathcal{D}) = \frac{P(\mathcal{D}|M)P(M)}{P(\mathcal{D})} \tag{31}$$

The marginal likelihood of model $P(\mathcal{D}|M)$ could be approximated using the particles at each iteration i (Wilkinson, 2011). Since the observation of particles only depends on the previous particle, We could therefore rewrite the term as,

$$P(\mathcal{D}|M) = P(\mathcal{D}_l) \prod_{l=1}^N P(\mathcal{D}_{i+1}|\mathcal{D}_{0:i}, M) \tag{32}$$

Assuming a priori equally likely models, the factor of $P(D)$ in (30) cancels between the two models and the Bayes Factors now becomes the ratio of two marginal likelihood,

$$B_{M1,M2} = \frac{P(\mathcal{D}|M_1)}{P(\mathcal{D}|M_2)} \tag{33}$$

Besides Bayes Factors, there also other methods commonly used to compare methods such as Akaike Information Criterion and Bayesian Information Criterion (Posada and Buckley, 2014; Bozdogan, 1987)

4 Result

In order to assess the performance of parameters inference and model comparison via particle filtering we created a test time series data set. The reaction model decribed in 3.3 was used as basis to generate the set with the aforementioned SSA of Gillespie. As the model describes a bistable switch the test data needs to contain sufficient information regarding both dominance states as well as transition phases between the states. To achieve this the simulation was organised into a cell lineage tree, starting with the simulation of single cell, which would then split into two daugthers cell. These cells inherit the protein levels of their parent and continue their simulation independently. With enough split phases the resulting time series tree can accurately represent the spectrum of possible system configurations. For our data set we choose 5 split phases resulting in 32 trajectories at the end of simulation. After generation of the test data the particle filter was applied with a particle count of 1000 and replicated two times to investigate the consistency of the parameter inference.

Parameter	Simulation	Prior	Replicate 1	Replicate 2	Replicate 3
k_1	120	100	104.489	105.533	104.497
k_2	120	100	104.585	104.771	104.099
k_3	0.4	0.3	0.356	0.351	0.344
k_4	0.4	0.3	0.355	0.337	0.356
K_{P1}	170	190	189.141	189.114	189.13
K_{P2}	170	190	188.209	189.295	189.606

Table 1. Exact value of the parameters used in the simulation to generate the test data and expected value of the prior used for particle filtering. Replicates show the expected value of the posterior distributions for each particle filtering run on the test data set.

Table 1 shows the results of the particle filtering test for each of the 6 paramters with the posterior distributions visualized in figure 4. The posterior distribution of the paramters k_3 and k_4 used the degradation reactions of `Pu.1` and `Gata1` respectivly, show the most recognisable motion towards their optimal values while the parameters k_1 and k_2 responsible for the production reactions show only little movement. No significant fitting of the inhibition paramters K_{P1} and K_{P1} can be inferred though all replicates show minimal movement. The replicates of the particle filtering are highly consistent in their results with expected values of their posterior distributions being marginally different. Furthermore the paramters k_3 and k_4 show a similar transformation of their particluar priors towards the shape of the resulting posteriors.

In order to investigate the capability of the particle filter for model comparsion a second model was created from the reaction model described in section 3.3. For this model the inhibitory kinetics of the production reactions were removed resulting in complete independence of protein levels between `Pu.1` and `Gata1`. The production parameters k_1 and k_2 were lowered from 120 to 90 to compensate for the removal of the inhibitory effect. Afterwards the particle filter algorithm was applied to test data set with 1000 particles for both the original reaction model with optimal paramters and the modified model without cross inhibiton. By tracking the particle weight of a particle trajectory at each given datapoint the particle filter can infer the likelihood for the Bayes Factors described in equation 33. For ease of computation the weights were transformed to log probabilities with base 2, resulting in the following Bayes Factor:

$$B_{M1,M2} = \frac{P(\mathcal{D}|M_{inhibitory})}{P(\mathcal{D}|M_{indepent})} = \frac{2^{-7379.655}}{2^{-7423.938}} \approx 2.14 \cdot 10^{13} \quad (34)$$

The inferred Bayes Factor shows a significantly increased likelihood of the inhibitory model used in the generation of test data set, compared to the indepent model, consequently identifying the correct model.

After testing the particle filter was then applied to the time series data set containing the cell lineage of a murine stem cells. 3 trees were chosen based on their protein distribution showing states of dominance for both transcription factor levels. With respect to the toggle switch hypothesis these trees sufficiently represent the spectrum of lineage choice possibilities. The particle filter algortihm was performed with 1000 particles for all 3 trees, each containing respectively 41, 25 and 42 living cells at the end of the experiment.

Parameter	Prior	Tree 1	Tree 2	Tree 3
k_1	800	799.369	798.596	800.094
k_2	400	398.35	401.315	400.126
k_3	0.01	0.0129	0.0125	0.0128
k_4	0.01	0.0185	0.0159	0.0223
K_{P1}	4000	4002.256	3993.389	4001.58
K_{P2}	15000	15002.5	14998.55	14993.86

Table 2. Expected value of the prior used for particle filtering for each parameter. Trees show the expected value of the posterior distributions for each particle filtering run on the respective cell lineage tree.

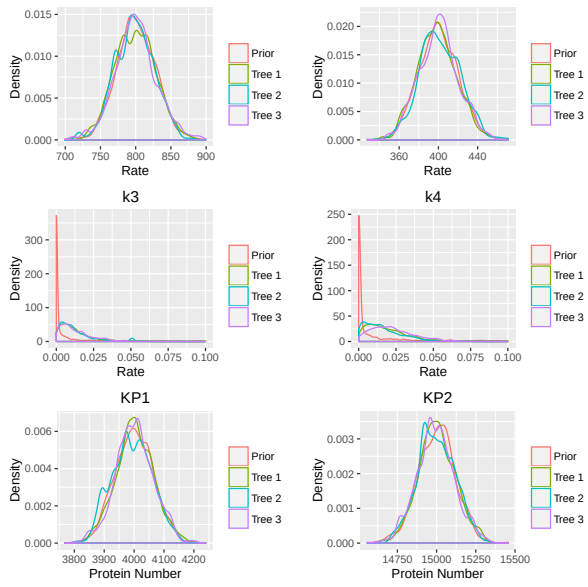


Fig. 5. Result of the particle filtering algorithm for each parameter of the reaction model on timer series data of murine stem cells. The red lines show the prior distribution for the parameters used in the particle filtering. Tree lines represent the posterior distribution of fitted paramters after particle filter run on the respective cell lineage tree.

Table 2 shows the fitted paramters derived from using particle filtering on the time series data of murine stem cells with the corresponding prior and posterior distribution shown in figure 5. The particle filter exhibits a similar paramter fitting performance compared to the test data set. On both the simulated as well as the experimental data the degradation rate parameters k_3 and k_4 show signifcant movement away from their expected prior values while the inhibitation paramters K_{P1} and K_{P1} show no

change of their expected value. However the production parameters k_1 and k_2 on experimential data set were not changed by the particle filtering.

After perfoming the particle filtering with the reaction model, the independent model used for model comparison on the test data set, was applied to experimental data an the following likelyhoods were derived:

$$B_{M1,M2} = \frac{P(\mathcal{D}|M_{inhibitory})}{P(\mathcal{D}|M_{indepent})} = \frac{2^{-2937.356}}{2^{-3640.912}} \approx 6.19{\cdot}10^{211} \quad (35)$$

As shown by the Bayes Factor the inhbitory model massively outperforms the independent model on the experimental data set.

5 Discussion

The test of the particle filtering algorithm on the simulated demonstrates the frameworks capability to infer model paramters based on tree structured time series data. However it shows uneven fitting performance given different paramter enviroments, with the degradation rates k_3 and k_4 being moved signficantly towards their optimum, while the production rate paramters k_1 and k_2 exhibit slower movement and the inhibitation parameters K_{P1} and K_{P1} were not changed at all. We suspect that this disparity of fitting performance is due to the difference in impact of the individual paramters on the protein levels and therefore the weight of the particles. The degredation reaction propensities depend on both the protein levels as well as the degredation rate and thus small changes of rates can have signifacnt influence on the protein levels. In comparison the production reaction propensities do not get scaled with protein numbers and hence the particle weighting is less susceptible to changes of production rates, resulting in a reduced fitting performance. As the inhibitation parameter only negativly scale the production reaction propensites, changes of these parameters shows even less impact on protein levels and the resulting particle weights. Another factor influencing the performance of the particle filter paramter inference is presence of equilibrium states in time series, during which the protein levels do not change by large margin. In these time frames the production and the degradation reaction propensities need to be evenly matched. This however can be achieved with any suboptimal production rate as long as the matching degredation rate can keep the equilibrium. In reaction model used during testing both production and degradtion rates lie below the optimal values, possibly resulting in paramter systm that can keep equilibrium stable. While the test data set also contains a multitude of transition phases with high alteration of protein levels, these phases are potentially not enought to allow for fitting of the parameters through particle filtering. This assumption is further amplified by remarkable consistency of the replicated particle filter applications on the test data.

The model comparison of the inhbitory and the indipendant model shows promising results as the Bayes Factor can accuratly discern which model was used for generate of the test data set by a large margin. This shows that although the independent model has the possibility to generate trajectories similiar to that of an inhibitory model, the resulting performance of the particle filtering is still negatively influenced.

The interpretation of the particle filtering on the experimental time series trees proves to be ambiguous. While the particle filter managed to significantly alter the degradation rates k_3 and k_4 with k_4 being more than doubled on Tree 3, the other parameters remain unchanged at the expected value of their respective priors. While this behavior is similar to the particle filter performance of the reaction model on the test data set, the results are insuffienct to derive whether the underlying mechanics can be represented through the model. As all particle filtering instances on the experimental data show to be highly consistent, the performance appears to be unaffected by potential disparity of the experimental data. One possible factor influencing the performance of the particle is the number of transition phases in the experimental data set. If the initial assumption of toggle switch deciding the cell fate holds true, then most of cells would spend large quantity of the experiment time in moderate equilibrium state until one transcription factor reaches a dominance threshold by chance, triggering a transition phase. As this transition represent the lineage choice, no futher transition phases can be observed in this branch of the cell lineage tree and protein levels are then governed by the underlying mechanism of MEP or GMP cells. This would severly limit the amount of transiton phases needed to infer parameters beyond maintaining the equilibrium.

Another possible explanation for particle filter performance is a discrepance between the model used for simulation and the underlying mechanics govering the transcription factors. The utilized reaction model is rather coarse as only contain production and degradation reactions. In comparison a more fine grained model containing the state of transcription sites corresponding to the proteins as well as mRNA levels, production, degradation and translation ratios could potentially better represent the stochastic process of the fundamental mechanics. Regarding the model representing the intial hypothesis of a cross inhibiton we mindful of a recent study on same experimental data set claiming independence of early myeloid lineage choice with regards to the transcription factors `Pu.1` and `Gata.1` (Hoppe *et al*, 2016). In their work the authors make sound arguements against a lineage choice governed by a stochastic process between `Pu.1` and `Gata.1`. If the authors suspicions holds true, then the rise of transcription factor production would be regulated by an external factor, majorly influencing the particle filters ability to infer production rate parameters during transition phases. However the model comparison between the indepent and the inhibitory model on the experimental data show a massive favor towards the inhibitory model, suggesting that cross talk between the transcription factors is definitively present. This is in accordance with the observation that at no time in the experiment the highest levels of both transcription could be observed simultaneously. Ultimately the described finding suggest the role of a cross inhibitory model to be the one of executing the lineage choiche rather than making it.

6 Conclusion

In the course of this project we managed to implement the particle filter algorithm as a framework for the inference of model parameters through the use of time series data. We demonstrate its ability to parameterize models with remarkable consistency and show the usefulness of particle trajectories for model comparison via Bayes Factors. While the particle filter did not manage to parameterize an agreeable model of our initial hypothesis, we found this to be in accordance with a recent study on the same data set, questining the role *Gata1* and *Pu.1* as toggle swtich deciding the cell fate of hematopoietic stem cells. This highlights our framework's capability to convey underlying mechanics prior unknown to us. This in turn hallmarks our framework potential as supporting tool for research scientists.

Acknowledgements

We would like to express our gratitude to Werner Mewes, Fabian Theis, Jan Quell and Anna Dieckman for facilitating the project. We would also like to thanks Philipp S. Hoppe, Michael Schwarzfischer, Dirk Loeffler, Konstantinos D. Kokkaliaris, Oliver Hilsenbeck, Nadine Moritz, Max Endele, Adam Filipczyk, Adriana Gambardella, Nouraiz Ahmed, Martin Etzrodt, Daniel L. Coutu, Michael A. Rieger, Bernhard Schauberger, Ingo Burtscher, Olga Ermakova, Antje Bürger, Heiko Lickert, Claus Nerlov

and Timm Schroder who agreed to provide us with the experiment data, without which we wouldn't be able to test our framework and models.

References

Feigelman, J. (2016). "Stochastic and deterministic methods for the analysis of Nanog dynamics in mouse embryonic stem cells." PhD Thesis, Technische Universität München, Munich, Germany.

Hoppe, P.S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K.D., Hilsenbeck, O., Mortz, N., ... & Etzrodt, M. (2016). Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature*, **535(7611)**, 299-302.

Graf, T., & Enver, T. (2009). Forcing cells to change lineages. *Nature*, **462(7273)**, 587-594.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, **297(5584)**, 1183-1186.

Orkin, S. H., & Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, **132(4)**, 631-644.

Filipczyk, A., Marr, C., Hastreiter, S., Feigelman, J., Schwarzfischer, M., Hoppe, P. S., ... & Hilsenbeck, O. (2015). Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nature cell biology*.

Zhang, J., Niu, C., Ye, L., Huang, H., He, X., Tong, W. G., ... & Harris, S. (2003). Identification of the haematopoietic stem cell niche and control of the niche size. *Nature*, **425(6960)**, 836-841.

Doucet, A., De Freitas, N., & Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice* (pp. 3-14). Springer New York.

Liu, J. S., & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443), 1032-1044.

Del Moral, P. (1996). Non-linear filtering: interacting particle resolution. *Markov processes and related fields*, **2(4)**, 555-581.

Del Moral, P., Doucet, A., & Jasra, A. (2012). On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, **18(1)**, 252-278.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, **81(25)**, 2340-2361.

Doob, J. L. (1945). Markoff chains–denumerable case. *Transactions of the American Mathematical Society*, **58(3)**, 455-473.

Chung, K. L. (1967). *Markov Chain*. Berlin: Springer-Verlag.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, **58**, 35-55.

Fahrmeir, L., Künstler, R., Pigeot, I., & Tutz, G. (2007). *Statistik: Der Weg zur Datenanalyse*. Springer-Verlag.

Wilkinson, D. J. (2011). Stochastic modelling for systems biology. CRC press.

Zechner, C., Pelet, S., Peter, M., & Koeppl, H. (2011, December). Recursive Bayesian estimation of stochastic rate constants from heterogeneous cell populations. In *2011 50th IEEE Conference on Decision and Control and European Control Conference* (pp. 5837-5843). IEEE.

Haseltine, E. L., & Rawlings, J. B. (2005). On the origins of approximations for stochastic chemical kinetics. *The Journal of chemical physics*, **123(16)**, 164115.

Sherlock, C., Golightly, A., & Gillespie, C. S. (2014). Bayesian inference for hybrid discrete-continuous stochastic kinetic models. *Inverse Problems*, **30(11)**, 114005.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, **53(5)**, 793-808.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52(3)**, 345-370.

Wilkinson, D. J. (2010). Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology. In *Proceedings of 9th Valencia International Meeting on Bayesian Statistics*. Oxford University Press (pp. 679-705).

Golightly, A., & Wilkinson, D. J. (2010). Markov chain Monte Carlo algorithms for SDE parameter estimation. *Learning and Inference for Computational Systems Biology*, 253-276.

Gonzalez, A., Uhlendorf, J., Schaul, J., Cinquemani, E., Batt, G., & Ferrari-Trecate, G. (2013). *Identification of biological models from single-cell data: a comparison between mixed-effects and moment-based inference* (Doctoral dissertation, INRIA).

Komorowski, M., Finkenstädt, B., Harper, C. V., & Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC bioinformatics*, **10(1)**, 1.

Michaelis, L., & Menten, M. L. (1913). Die kinetik der invertinwirkung. *Biochem. z*, **49(333-369)**, 352.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.

Nodelman, U., Allen, C., & Perry, J. (2003). *Stanford encyclopedia of philosophy*.

Cao, Y., Gillespie, D. T., & Petzold, L. R. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, **124(4)**, 044109.

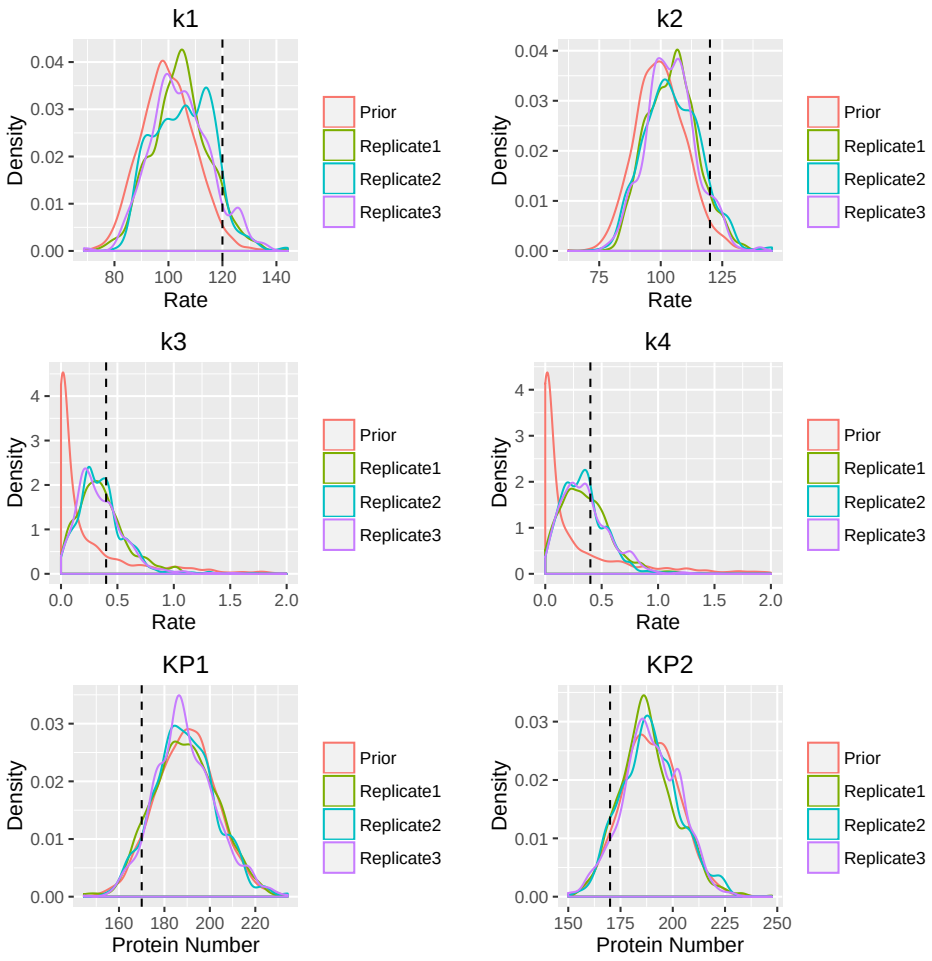


Fig. 4. Result of the particle filtering algorithm for each parameter of the reaction model on the test data set. The dashed lines indictated the optimal parameter values used in the data creation. The red lines show the prior distribution for the parameters used in the particle filtering. Replicate lines represent the posterior distribution of fitted paramters after particle filter run.