Bioinformatics, 31, 2015, i89–i96
doi: 10.1093/bioinformatics/btv257
ISMB/ECCB 2015

OXFORD

Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data

Andrea Ocone[1], Laleh Haghverdi[1], Nikola S. Mueller[1] and Fabian J. Theis[1,2,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany and [2]Department of Mathematics, Technical University Munich, 85747 Garching, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: High-dimensional single-cell snapshot data are becoming widespread in the systems biology community, as a mean to understand biological processes at the cellular level. However, as temporal information is lost with such data, mathematical models have been limited to capture only static features of the underlying cellular mechanisms.

Results: Here, we present a modular framework which allows to recover the temporal behaviour from single-cell snapshot data and reverse engineer the dynamics of gene expression. The framework combines a dimensionality reduction method with a cell time-ordering algorithm to generate pseudo time-series observations. These are in turn used to learn transcriptional ODE models and do model selection on structural network features. We apply it on synthetic data and then on real hematopoietic stem cells data, to reconstruct gene expression dynamics during differentiation pathways and infer the structure of a key gene regulatory network.

Availability and implementation: C++ and Matlab code available at https://www.helmholtz-muenchen.de/fileadmin/ICB/software/inferenceSnapshot.zip.

Contact: fabian.theis@helmholtz-muenchen.de

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

1 Introduction

Unraveling the dynamics of gene regulatory programs is a challenging problem, due to multitude of variables and mechanisms involved in the gene control machinery. As genes represent fundamental units in regulatory networks, the problem is commonly restricted to understand how they interact and express themselves, to generate specific biological functions. This gave rise to a variety of mathematical methods which, combined with experimental measurements, provided a way to reconstruct gene regulatory processes and estimate unknown kinetic parameters. Different classes of approaches have been proposed, from ODE-based models (Honkela *et al.*, 2010; Sanguinetti *et al.*, 2009) to stochastic models (Wilkinson, 2009), and a number of optimization techniques have been developed for inference and parameter estimation in gene regulatory models (Liepe *et al.*, 2014; Ocone *et al.*, 2013; Stathopoulos and Girolami, 2012).

To learn network dynamics, all these models require temporal data sources, such as mRNA or protein time-courses. However, as recent advances in single-cell technologies offer the possibility to analyse gene expression simultaneously in hundreds of individual cells (Citri *et al.*, 2012), time-course data are not always available for massive single-cell dataset. This prevents from using elaborated methods developed in the last decades on data coming from newest technologies such as single-cell RNA sequencing or high-throughput quantitative polymerase chain reaction.

Typically, data from such technologies are given in form of single-cell snapshot data. Broadly, single-cell snapshot data consist of gene expression measurements collected from multiple single cells at a single time point. So defined, they represent static data which reflect single-cell states at a single time point. However, as cells may have different stochastic behaviours during the same process (Elowitz *et al.*, 2002), a sort of temporal information is still retained in snapshot data.

Here, we present a framework which allows to extract gene regulatory dynamics directly from single-cell snapshot data. The framework allows reconstruction of individual network nodes' dynamics, estimation of kinetic parameters and computation of Bayes' factors to determine how network nodes functionally interact.

The combination of dimensionality reduction with a clustering method and an efficient algorithm to order single cells by time provides a way to reconstruct gene expression pseudo time courses from different cellular processes. These are used to compare different ODE-based transcriptional models and select the one which best

explains the data. As transcriptional models incorporate structural knowledge in form of Boolean logic gates and presence/absence of regulatory edges, the model selection step represents indeed a way to refine the gene network structure.

In contrast to recently developed frameworks (Bendall *et al.*, 2014; Trapnell *et al.*, 2014), we show how dynamics extracted from single-cell snapshot data can be used to recover quantitative aspects about the underlying regulatory network. We first test our framework on two simulated datasets. The first consists of a feed-forward loop (FFL) network motif, which exhibits the behaviour of a pulse generator. The other dataset is generated by a more complex regulatory network, including toggle switches, which mimics the dynamics of cell differentiation process.

Finally, we present an application on hematopoietic stem cells (HSCs) data generated by a high-throughput single-cell quantitative polymerase chain reaction experiment during cell differentiation (Moignard *et al.*, 2013). We reconstruct both dynamics and structure of a key gene regulatory network (GRN) and show that our framework provides biological insights not accessible through standard *in silico* analyses.

2 Modular framework for learning GRN dynamics

Our approach is based on a modular framework, depicted as a block diagram in Figure 1. Gene expression single-cell snapshot data represent the input to the framework (Fig. 1A). The outcome of the analysis is 2-fold: we estimate kinetic parameters for transcriptional dynamics (Fig. 1H) and we refine the GRN topology (Fig. 1G). Inferred GRN structure and parameters can eventually be used to generate model predictions in presence of structural and/or dynamical perturbations.

Each module in the framework performs a crucial function, which is summarized below and described in detail in Supplementary Information. The first module performs dimensionality reduction (Fig. 1B), which allows to embed high-dimensional snapshot data into a low-dimensional space. Here, we use a nonlinear method, known as diffusion map (Coifman *et al.*, 2005), which represents the probabilistic version of spectral clustering techniques (Nadler *et al.*, 2005). By means of diffusion map, a large number of cells can be embedded and visualized in a low-dimensional (i.e. 2D–3D) space, according to their multivariate gene expression value. In this low-dimensional space, similarity between cells is encoded by their Euclidean distance. In particular, by applying an *ad hoc* clustering method on the cells in low-dimensional space (Fig. 1C), we are able to separate a number of branches associated with different cellular processes (i.e. differentiation pathways). The number and selection of branches depend on a user-defined initial cell, therefore an approximate position of this cell is required as prior information. Although this is generally not available, a prior knowledge of expression profile of key genes combined with visual inspection in diffusion map embedding is generally sufficient to locate an approximate initial cell.

As gene expression is a function of time, cells can potentially also be ordered by time. We use a recently developed algorithm, Wanderlust (Bendall *et al.*, 2014), to order single cells along discrete paths, in branches identified in the embedded data space. These paths do not represent real time but rather a pseudo time variable, which depends on the intrinsic cellular process. As a result, we reconstruct gene expression dynamics of different genes in form of pseudo time-series (Fig. 1D).

As Wanderlust works only for non-branching processes, the *ad hoc* clustering method is necessary to separate multiple branches before ordering the cells by time. Note that Wanderlust is applied directly on high-dimensional data; on the other hand, performance of the clustering method in low-dimensional diffusion map embedding is better than directly in high-dimensional data (see details in Supplementary Information).

To learn gene regulatory dynamics from pseudo time-series data, we describe gene expression dynamics of individual network nodes using mathematical models. The choice of right level of model abstraction is crucial to achieve a certain task. Our application requires a mathematical model, which is flexible enough to explain nonlinear gene expression dynamics and allows for an efficient and

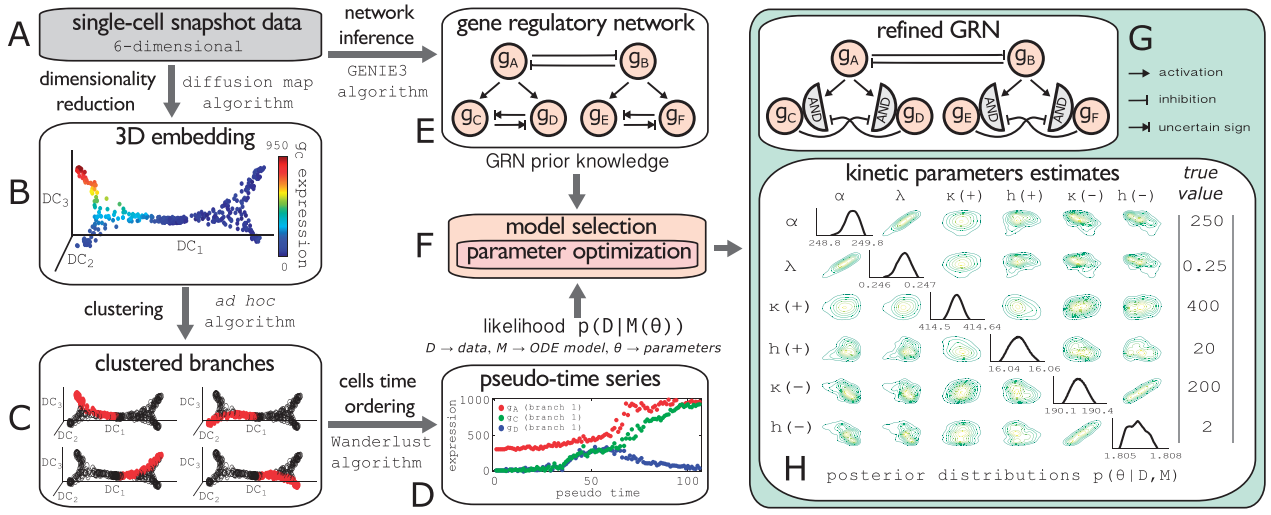

Fig. 1. Diagram of the framework. High-dimensional single-cell snapshot data (static data) are used as input in two paths (**A**). The first is a combination of a dimensionality reduction method [i.e. diffusion map algorithm (Coifman *et al.*, 2005)] (**B**), clustering (**C**) and cell time-ordering [i.e. Wanderlust algorithm (Bendall *et al.*, 2014)], which provides pseudo time-series (dynamic data) (**D**). Axis labels DC in the low-dimensional space, represent the first few diffusion components, i.e. eigenvectors. The second is a network inference algorithm [i.e. GENIE3 (Huynh-Thu *et al.*, 2010)] generating a coarse GRN structure (**E**), which is used as prior knowledge during model selection. A likelihood function, which links transcriptional models *M* to the pseudo time-series data *D*, is used in a Bayesian framework to perform model selection and parameter estimation (**F**). Output is represented by a refined GRN structure (**G**) with corresponding posterior estimates of kinetic parameters (**H**)

accurate solution of the parameter estimation problem. To this aim, we use the following ODE-based model to describe gene–gene interaction:

$$\dot{y}(t) = \alpha f(x(t),\theta) - \lambda y\,,$$
$$f(x(t),\theta) = \begin{cases} \dfrac{x^h}{x^h+\kappa^h}, & \text{if } x \text{ is activating} \\ \dfrac{\kappa^h}{x^h+\kappa^h}, & \text{if } x \text{ is inhibiting}\,, \end{cases} \quad (1)$$

where x and y represent mRNA concentrations of input and target gene, respectively. Kinetic parameter α represents production rate and λ decay rate of target gene expression; $\theta \equiv (\kappa, h)$ are parameters of a nonlinear Hill-type function $f(x(t),\theta)$ (Hao and O'Shea, 2011), where κ and h are dissociation constant and Hill coefficient, respectively. [We assume that mRNA concentration y of target gene can be used as proxy for concentration level of its active transcription factor (TF). This is valid by considering that post-transcriptional modifications occur on a faster timescale with respect to transcription and translation process. As a direct consequence, kinetics parameters in Equation (1) will take into account of both transcription and translation mechanisms. Alternatively, statistical models where protein states are treated as latent variables could be adopted (Ocone *et al.*, 2013).]

Generally, the expression of a target gene y is regulated by the activity of a number M of inputs x_i (with $i = 1, \ldots, M$), which can be combined according to different logical expressions. Therefore, a wide range of possible models can be used to describe interactions between M inputs on target y. The following models

$$\dot{y}(t) = \alpha \prod_{m=1}^{M} f_m(x_m(t),\theta_m) - \lambda y\,, \quad (2)$$
$$\dot{y}(t) = \alpha \sum_{m=1}^{M} f_m(x_m(t),\theta_m) - \lambda y\,, \quad (3)$$

encode different logical expressions to combine M input genes, respectively AND and OR logic gates. Generalization to mixtures of different logical expressions is straightforward.

For parameter estimation, we use an approximation method based on Markov chain Monte Carlo (MCMC). In each MCMC iteration, a Gaussian likelihood is computed using the solution y of model ODEs at observation pseudo times. As we are considering a network of interacting genes, likelihood computation turns to be a recursive system, which we solve by using Gaussian process emulators for input gene functions $x_m(t)$ (O'Hagan, 2006). In other words, parameter estimation of a GRN with N genes is decomposed in N different optimization problems correspondent to N subnetworks. Each subnetwork is composed of a single target gene, which is regulated by a number of input genes. Given emulators for its inputs, each subnetwork is *conditionally independent* on the others (Georgoulas *et al.*, 2012). As the dimensionality of parameter space is reduced in each subnetwork, this strategy turns to be very efficient and scalable with GRN size.

To select which ODE model explains better pseudo time-series data, parameter optimization is integrated with a model selection step. In addition to more simple statistics as Akaike (AIC) and Bayesian information criterions (BIC), we do Bayesian model comparison by computing Bayes' factors through thermodynamic integration (Calderhead and Girolami, 2009).

As the number of ODE models to compare grows combinatorially with the number of genes, the amount of models is thinned out by using *a priori* knowledge on GRN structure. We adopt GENIE3 (Huynh-Thu *et al.*, 2010), an efficient method based on random forest, to obtain a coarse network structure, which we then refine through model selection (Fig. 1E).

As the framework represents a combination of multiple methods and each method depends on user-provided parameters, it is relevant to evaluate how its global performance depends on these parameters. For this reason, we perform multiple tests on simulated data to validate robustness of results to different parameter choices (see Supplementary Information).

In the following sections, before applying our framework on real data, we assess its performance on two synthetic datasets.

3 Simulated data: FFL network motifs

FFLs are three-gene networks which have been of large interest in the recent literature (Alon, 2006; Ocone and Sanguinetti, 2011). They are composed by a master TF g_X, regulating a slave gene g_Y and a target gene g_Z. The target gene is in turn regulated by the activity of the slave TF g_Y. As TFs can activate or inhibit their targets, there can be eight possible FFL types according to the sign of the three edges. Here, we focus on the incoherent type-I (I1) FFL (Fig. 2B), where the master TF activates both slave g_Y and target gene g_Z, so that a positive activity of g_X determines an increasing expression for g_Y and g_Z. However, as the target gene is repressed by the slave TF g_Y, its expression starts decreasing to steady state level as soon as g_Y expression is greater than a given threshold (which depends on the parameters of g_Z activation function). As a consequence, g_Z expression exhibits a pulse behaviour, which accelerates the gene expression dynamics (Fig. 2D). This feature is so fundamental for accurate control of cellular processes that I1-FFL is one of the two most common FFL types found in GRNs of *Escherichia coli* and yeast (Mangan *et al.*, 2006).

We simulate single-cell snapshot data sets using an I1-FFL where input genes g_X and g_Y interact on the common target g_Z through an OR logic gate. Stochastic realizations for $N = 300$ cells are generated using Euler–Maruyama method and expression values for the three

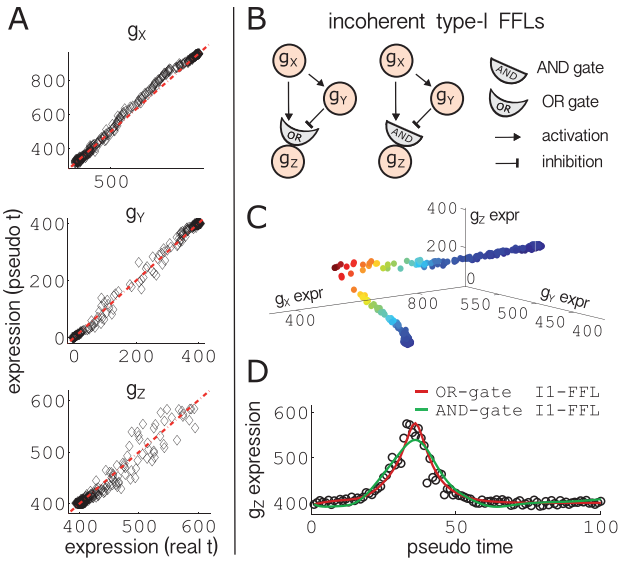

Fig. 2. (**A**) Comparison of reconstructed pseudo tim-series with time-series generated from true model. (**B**) Structure of incoherent type-I FFLs. (**C**) Three-dimensional embedding of snapshot data generated from OR-gate I1-FFL. Colours encode g_Z expression levels. (**D**) Fit of AND-gate (solid green) and OR-gate (solid red) I1-FFL models to pseudo time-series data (black circles)

genes at a random time point are used to create the single-cell snapshot data.

Dimensionality reduction is not necessary for FFLs, as cells can be embedded in a 3D space where each dimension represents a gene (Fig. 2C). Application of Wanderlust algorithm in this case is straightforward, as cells are distributed along a single non-branching path. This is essentially due to the fact that FFLs are not bistable systems like the ones we will discuss later on, but they always evolve according to a characteristic dynamics.

To quantify how well cell time-ordering performs, we compare expression values from pseudo time-series to expression values obtained from deterministic FFL model simulation. Results in Figure 2A show a good agreement between pseudo time and real time; small discrepancies are due to the fact that we compare stochastic data with a deterministic simulation where intrinsic noise is null. Gene g_Z shows slightly larger discrepancies compared with g_X and g_Y, as its dynamics is more complex (i.e. overshoot dynamics instead of monotonic functions) and time-ordering is less efficient.

Once a rough GRN structure is obtained through network inference, we aim to refine this structure. This is done by learning presence/absence of regulatory edges; uncertain regulatory edge signs (i.e. activation/inhibition); logical interaction occurring between master and slave TFs on target gene promoter. By using transcriptional models with single/multiple inputs, with AND-gate and OR-gate (Eq. 2 and 3), we estimate kinetic parameters and do model selection.

Model selection results are reported in Supplementary Information. For models with highest AIC/BIC values, marginal likelihoods have also been computed for Bayesian model comparison. Bayes' factor between OR-gate and AND-gate FFL results $R_{\mathrm{OR/AND}} > 100$. According to Jeffrey's interpretation (Jeffreys, 1961), a Bayes' factor $R_{\mathrm{OR/AND}} > 100$ clearly shows that transcriptional model with OR-gate can explain pseudo time-series reconstructed from OR-gate I1-FFL snapshot data decisively better than competitive AND-gate model. Results are validated by visual inspection through fitting with optimized parameters (Fig. 2D).

The goodness of fitting for favoured models is also reflected in estimated parameter values. Parameter posterior distributions are reported in Figure 3, together with true parameters used to simulate snapshot data. Average of relative errors, computed using maximum *a posteriori* estimates, is around 17%. However, parameter estimation performance is mixed: mode of posterior distributions for some parameters is very close to true values, whereas for other parameters, such as Hill coefficients h, it is not as good. Variations of Hill coefficients h determine only small changes in the slope of g_Z expression overshoot, therefore even small error sources in gene expression values or pseudo time reconstruction make estimation of h nontrivial. Plots of bivariate marginal distributions (Fig. 3, left) show high correlation between parameters α and λ. Although we are not able to recover their true values, nonetheless the ratio between their posterior estimates is exactly the true one (Fig. 3, right). The problem can be reduced by placing an informative Gaussian prior over decay rate parameter λ; in this case parameter estimation improves and reduce the average of relative errors to 14.5%.

To assess predictive power of our framework, we have used inferred GRN structure and estimated parameters to generate predictions under different perturbed conditions. Perturbations can be in the form of structural changes of the network and/or in the form of dynamical changes, e.g. variation of kinetic parameters values. Furthermore, single perturbations can be combined to obtain simultaneous perturbations. Here, we consider three perturbations: deletion of slave gene g_Y, gene knockdown to achieve 3-fold increase in

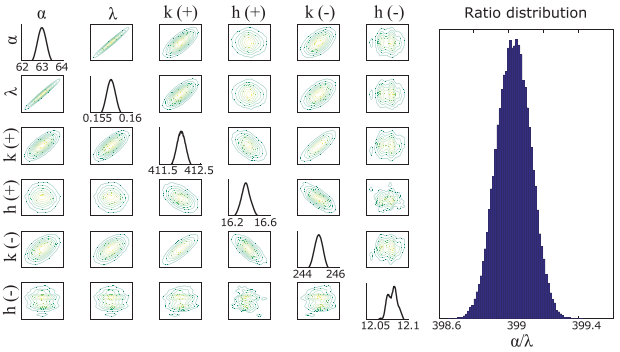

Fig. 3. Left: posterior parameter distributions, represented as univariate and bivariate marginal distributions, using OR-gate I1-FFL. True parameters used to simulate data are the following: $\alpha = 100$, $\lambda = 0.25$, $\kappa_+ = 400$, $h_+ = 20$, $\kappa_- = 200$, $h_- = 10$. Right: ratio distribution obtained from posterior distributions over α and λ. True parameters ratio is given by $\alpha/\lambda = 400$

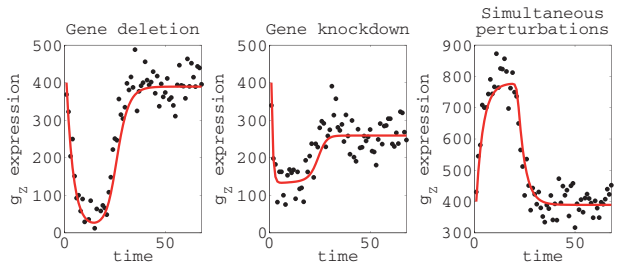

Fig. 4. Predictions generated by GRN inferred from snapshot data, under three different perturbation conditions. Noisy data generated from perturbed true OR-gate I1-FFL system (black dots), compared with predictions generated using inferred GRN and estimated parameters (solid red lines)

degradation rate λ; simultaneous change of dissociation rate κ_+ (10x decrease) and κ_- (5x decrease). Time-series were simulated using the true system under each perturbation and compared with dynamics generated (under same perturbations) by inferred GRN and estimated parameters. A detailed procedure is reported in Supplementary Information. Predicted dynamics under perturbation conditions fit very well to observations from true perturbed system (Fig. 4), proving that dynamic information extracted from snapshot data of FFLs can be used to generate useful predictions.

Our results show that, using high-dimensional single-cell snapshot data, our framework provides a way to reconstruct gene expression dynamics and learn logic gates in FFL network motifs. FFLs represent a simple example; in the next section we demonstrate the power of our framework on a more complex regulatory network.

4 Simulated data: toggle switch network

A toggle switch is a double-negative loop motif composed by two genes repressing each other (Gardner *et al.*, 2000). Because of its robust bistable dynamics, this motif is useful in systems such as the lysis-lysogeny switch in λ-phage (Ptashne and Gann, 2002) and it is believed to be actively present in GRNs regulating stem cells differentiation (Wang *et al.*, 2009).

Here we consider a six-gene network which mimics the process of decision making during stem cells differentiation (Fig. 1G). The network includes three toggle switches in a symmetrical hierarchical structure: toggle switch $g_A - g_B$, at an early differentiation stage, controls the state of two toggle switches $g_C - g_D$ and $g_E - g_F$, at a

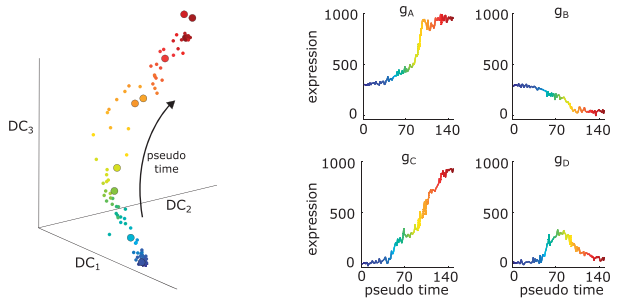

Fig. 5. Left: application of Wanderlust algorithm on one diffusion map's branch from toggle switch simulated snapshot data. Colours encode time progression along the resulting trajectory. Right: reconstructed pseudo time-series, obtained by plotting gene expression values in time-ordered cells

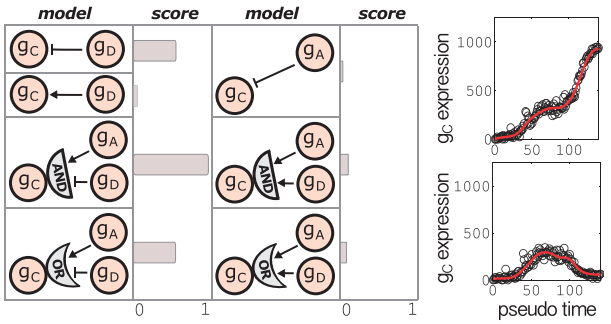

Fig. 6. Left: model selection results for target g_C. Model score in interval $[0, 1]$ represents a normalized AIC value: score $= (\text{AIC} - m)/\max(\text{AIC} - m)$, where m represents minimum AIC among all models. Right: fit of selected model (solid red) to g_C pseudo time-series data reconstructed from two different branches (black circles)

late differentiation stage. Expression of late stage genes, g_C, g_D, g_E and g_F, reflects the commitment to four different cell fates. For each differentiation path, only a single late stage gene is expressed, depending on the state of upstream toggle switches. For example, when toggle switch $g_A - g_B$ activates g_A transcription, then g_A and g_B expressions will increase and decrease, respectively. As a consequence, genes g_C and g_D will both start to be transcribed, until toggle switch $g_C - g_D$ will keep activating only one gene and repressing the other one. This mechanism of expressing only g_C (or g_D) works because edges on late stage genes are combined through logic AND gates.

We again reconstruct GRN dynamics and refine its structure, by using simulated single-cell snapshot data and prior GRN structural information obtained though network inference. Snapshot data are simulated with 400 cells and same kinetic parameters for all three toggle switches. Application of diffusion map algorithm and selection of first three eigenvectors produces an embedding where four branches correspondent to the different cell fates can be clearly identified (Fig. 1B). Pseudo time-series are finally reconstructed for each branch, through Wanderlust algorithm (Fig. 5).

As kinetic parameters are the same for each toggle switch, we focus only on a single gene, g_C, and learn how inputs genes (i.e. g_A and g_D), are combined on g_C activation function. We compare all possible transcriptional models, with single or double inputs and with different logic gates (Fig. 6, left). During model selection, parameters are optimized by using pseudo time-series from branches where g_C expression exhibits a dynamic behaviour (Fig. 6, right), i.e. two branches where g_A has high expression levels. Model selection results (Fig. 6, left) show that the transcriptional model with activating input g_A and inhibiting input g_D, combined through AND gate, is clearly favoured with respect to all the other models. Analogous results have been obtained for gene g_D (reported in Supplementary Information).

Parameter estimation is performed using simultaneously g_C and g_D pseudo time-series, with transcriptional AND-gate models for both g_C and g_D activation function. Means of parameter posterior distributions (Fig. 1H) reflect true parameter values (average of relative errors is around 6.8%), except again for Hill coefficients. By using more than a single branch for the optimization, the problem of correlation between parameters α and λ is reduced. To improve the accuracy of parameter estimation, our approach should be combined with a specific experimental design (Silk *et al.*, 2014), but this is out of the scope of this work.

As mentioned above, these results have been obtained using pseudo time-series from two (out of four) branches. By using all four branches, the estimation accuracy should further increase, but that is nontrivial, as gene g_C expression in the remaining two branches is just noise.

Another way to improve parameter estimation would be to consider a time-shift error produced by the time-ordering algorithm. An adjusted time-ordering method would require a study on the diffusion maps topology, which will be deferred to a further work. However, the accuracy of pseudo time-series reconstruction performed here is adequate for both model selection and reconstruction of transcriptional dynamics.

To assess the robustness of the framework, we performed analyses on different simulated dataset obtained by changing both observation noise level and noise model. Results show that parameter estimation accuracy is consistent over the different datasets (full details in Supplementary Information).

Here, we showed how our framework is able to get rid of bimodal gene expression (due to presence of multiple toggle switches) and recover dynamics of a more complex 6-gene GRN. As in the FFL case, by using reconstructed pseudo time-series and ODE-based transcriptional models, we were also able to select correct logics in the transcriptional activation function.

5 Logic learning in a hematopoietic GRN

Understanding stem cell differentiation is fundamental for advances in cell reprogramming research, with potentially practical applications in regenerative medicine and drug development (Cherry and Daley, 2012; Inoue and Yamanaka, 2011). In particular, mechanisms of cellular decision making are regulated by multiple dynamics of underlying GRNs. A paradigm for studying how these dynamics give rise to different cells fates is represented by blood cell formation process, known as hematopoiesis (Orkin and Zon, 2008).

Recently, Moignard *et al.* (2013) analyzed the expression of 18 key genes in 597 single cells isolated from mouse bone marrow. Cells were belonging to multiple stages during differentiation (Fig. 7A), from HSCs to four different progenitor cell populations: pre-megakaryocyte/erythroid progenitors (PreMegE), lymphoid-primed multipotent progenitors (LMPP), granulocyte-macrophage progenitors (GMP) and common lymphoid progenitor (CLP). They clustered cells according to expression states characteristic of each cell population and performed correlation analysis to infer GRN regulatory links. In particular, they predicted strong correlations between TFs GATA2, GFI1 and GFI1B, which were then validated using transcription and transgenic assays.

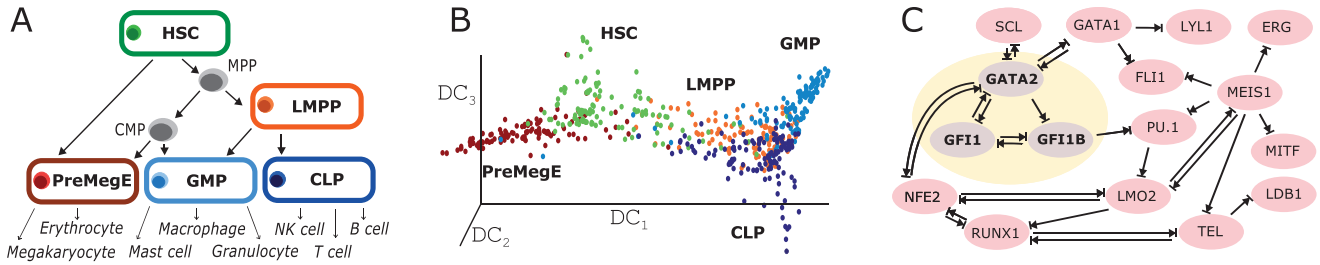

Fig. 7. Hematopoietic data. (**A**) Differentiation lineages of HSCs. (**B**) Diffusion map of hematopoiesis single-cell snapshot data set. (**C**) GRN obtained after network inference on snapshot data

Here, we focus on this regulatory triad (GATA2, GFI1 and GFI1B) and refine the subnetwork underpinning its dynamics during differentiation.

Figure 7B shows the embedding of single cells after dimensionality reduction with diffusion map algorithm and selection of first, second and third eigenvectors (i.e. diffusion components). Colours indicate different cell types, identified using cell surface markers (Cell surface markers are also used to separate branches, without need of a clustering method.). Two features can be observed: first, diffusion map is able to separate different cell types; second, the 3D embedding consists of three noisy but well-defined branches. Although it may wrongly seem that four branches are present in the embedding, in reality the fourth branch is composed of HSC cells and represents the starting point of the process. The three branches clearly show the three cell differentiation pathways as reported in Figure 7A: HSC → PreMegE, HSC → LMPP → GMP and HSC → LMPP → CLP. For sake of clarity, in the following we refer to these pathways as PreMegE, GMP and CLP, respectively. Among these pathways, we consider only pathways GMP and CLP, since reconstructed GFI1 pseudo time-series exhibit no dynamical behaviour during differentiation to PreMegE (Supplementary Fig. S19).

Network inference through GENIE3 algorithm and correlation analysis generate GRN in Figure 7C, where TFs of interest are highlighted by a yellow circle. For each of these TFs, we aim to compare different ODE models by combining different features: presence of single/multiple inputs on target gene; positive/negative edge sign and logical interactions (AND/OR gates and their combinations). A comprehensive list of all tested models is reported in Supplementary Information.

Results for target GATA2 have been obtained using pseudo time-series from pathway CLP. In fact, in dynamics reconstructed using pathway GMP, SCL exhibits residual bimodal expression and NFE2 a nearly flat profile. On the other hand, expression of GATA1 remains always at very low values in both pseudo time-series from CLP and GMP, therefore we do not take it into account.

Model selection favours two inputs influencing GATA2 expression: the first is a direct activation by SCL and the second is a direct inhibition by GFI1. Results are promising, as these interactions have been known or experimentally validated (Moignard *et al.*, 2013). On the other hand, a third predicted input, given by logical AND gate between GFI1 and NFE2, opens new questions to be investigated in laboratory.

Model selection for targets GFI1 and GFI1B have been performed using only pseudo time-series from pathway GMP, as input GFI1B exhibits bimodal expression in pathway CLP.

Results for target GFI1 show that only inhibitor GFI1B influences GFI1 expression. This interaction is part of a toggle switch between GFI1 and GFI1B (Moignard *et al.*, 2013).

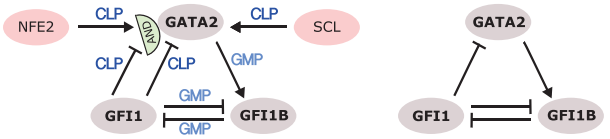

Fig. 8. Refined subnetwork using reconstructed pseudo time-series data (left) compared with experimental validated network in (Moignard *et al.*, 2013) (right). Text label on each edge indicates the differentiation pathway used to predict that edge. In the experimental validated network, the known relationship between SCL and GATA2 (Moignard *et al.*, 2013) is not shown; on the other hand, NFE2 is only predicted to be positively correlated with GATA2 in (Moignard *et al.*, 2013)

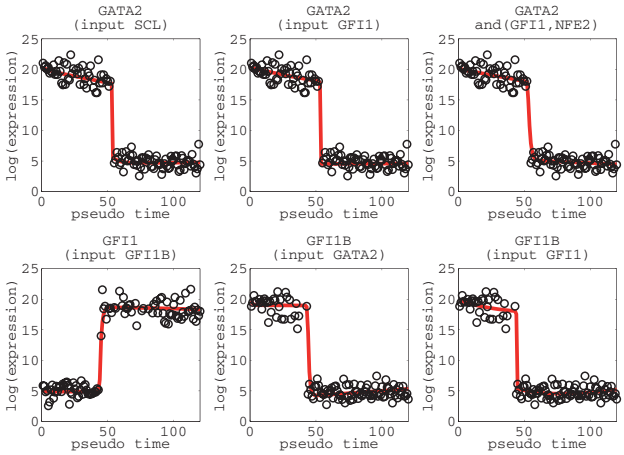

Fig. 9. Fitting of pseudo time-series from hematopoietic snapshot data with selected models. Observations from pseudo time-series (black circles), compared with model fitting (red lines)

The structure of this toggle switch is fully predicted, since also model selection for target GFI1B favours negative regulation of GFI1B by GFI1. Furthermore, results for target GFI1B predict also positive regulation of GFI1B by GATA2, so that GATA2 carries out a modulating function for the toggle switch, as experimentally validated in (Moignard *et al.*, 2013). However, a combinatorial model involving both GFI1 and GATA2 to regulate GFI1B is not predicted. Possibly, toggle switch GFI1-GFI1B represents such a central role in hematopoietic differentiation, that influence of a third regulator (i.e. GATA2) is included through a function more complex than AND/OR logic gates.

Figure 8 (left) shows a summary of the refined subnetwork: for each TF, selected edges are reported with corresponding data set from which results have been obtained. Fitting results are showed in Figure 9 and full details of transcriptional models are reported in Supplementary Information.

In summary, using dynamical information from GMP and CLP pathways, we are able to reconstruct gene expression dynamics during haematopoiesis. Pseudo time-series show how TFs behave during differentiation along pathways GMP and CLP. This reflects knowledge about gene expression of key TFs: for example GATA2 is known to be expressed at high levels in HSCs but at lower levels in LMPPs and GMPs. Furthermore, we are also able to refine the key subnetwork for the regulatory triad GATA2, GFI1 and GFI1B. Regulatory edges were not only evaluated in terms of sign but also directionality. On the other hand, directed edges were not predicted in Moignard *et al.* (2013) but only experimentally validated. Our framework also provides new testable hypothesis regarding interactions of TFs on target genes promoters, as in the case of TFs NFE2 and GFI1 for target GATA2.

6 Discussion

As mathematical modelling is becoming more and more crucial in supporting experimental research in systems biology, computational frameworks need to constantly evolve and adapt to new technologies. In this context, recent advances in high-dimensional single-cell technologies are providing massive amount of data in form of single-cell snapshot data. Although a number of statistical tools have been developed to model gene expression in snapshot data, they were only able to provide limited insights into biological mechanisms. In particular, as snapshot data do not directly represent how concentration of molecular species changes over time, present methods are only able to extract static information from data.

Here, we have presented a framework which allows to reconstruct dynamics of gene network nodes from single-cell gene expression snapshot data. By combining a dimensionality reduction method with an algorithm to time-order single cells, we are able to reconstruct pseudo time-series which provide a good approximation to time-series simulated from the real system. Through an efficient optimization strategy, pseudo time-series are then used to calibrate nonlinear ODE models which explain complex transcriptional dynamics. Furthermore, by computing Bayes' factors, we are also able to refine GRNs structures by selecting among transcriptional models with different structural information, i.e. different logic gates in activation functions.

The framework is modular and requires a number of sequential steps, which are independent on each other. This has the advantage that each step can be independently improved, without reducing the efficiency of the entire workflow. Many of these steps, especially dimensionality reduction techniques and parameter estimation methods, have already been studied in great detail. However, to the best of our knowledge, similar comprehensive frameworks have not been yet developed: our idea represents the first example of quantitative modelling GRNs dynamics without using protein or gene expression time-courses but only high-dimensional single-cell snapshot data. Only recently, some approaches have been developed to analyse high-dimensional biological data (Amir *et al.*, 2013) and their underlying dynamics (Amat *et al.*, 2014; Bendall *et al.*., 2014; Trapnell *et al.*., 2014) but purely for a visualization purpose or qualitative analyses.

Analysis on multiple simulated data sets demonstrates that we are able to infer kinetic parameters describing the underlying dynamics and recover logic gates in activation functions. Results on single-cell snapshot data obtained during blood stem cell differentiation show that we can reconstruct gene expression dynamics along differentiation pathways. Compared with competing strategies, we can not only determine correlations among genes in the GRN but also predict edge directions. In particular, GATA2 was predicted to be inhibited by GFI1 and to activate GFI1B, thus possibly working as a modulator of the known toggle switch between GFI1 and GFI1B. Gene GATA2 has been found to be involved in myeloid leukaemia (Hahn *et al.*, 2011); a better understanding of its role will not only help to get more insights into cell decision making but also to gain more knowledge about cancer pathologies.

From the computational point of view, our work opens possibilities to several further developments. Although diffusion maps have been proved to resolve nonlinear differences in single-cell expression data (Moignard *et al.*, 2015), an improvement in the time-ordering algorithm would substantially reduce the error in pseudo time-series. In addition, at the statistical modelling level, we have considered ODE-based transcriptional models without taking into account of intrinsic stochasticity of gene expression. Single cells provide more information with respect to cell populations, but such information could be retrieved only by using stochastic models of gene expression. On the other hand, parameter estimation and model selection with stochastic models are challenging and will require more advanced methods compared with those used in this work (Liepe *et al.*, 2014).

Despite we restricted the analysis to small/medium size GRNs, we believe that our work may evolve to the analysis of larger network and with more complex dynamics, e.g. including feedback loops. A greedy-type strategy could be adopted in model selection to cope with a larger number of dynamical models. However, an integration of structure inference and dynamics modelling would remain challenging in the case of larger networks (Oates *et al.*, 2014).

Acknowledgement

We gratefully acknowledge useful discussions with Guido Sanguinetti and Jan Hasenauer.

Funding

This work was supported by an ERC starting grant award (LatentCauses) [to F.J.T.].

Conflict of Interest: none declared.

References

Alon,U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC Press, London.

Amat,F. *et al.* (2014) Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods*, **11**, 951–958.

Amir,E.D. *et al.* (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–552.

Bendall,S.C. *et al.* (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, **157**, 714–725.

Calderhead,B. and Girolami,M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data Anal.*, **53**, 4028–4045.

Cherry,A.B. and Daley,G.Q. (2012) Model selection in systems biology depends on experimental design. *Cell*, **148**, 1110–1122.

Citri,A. *et al.* (2012) Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.*, **7**, 118–127.

Coifman,R.R. *et al.* (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA*, **102**, 7426–7431.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Gardner,T.S. *et al*. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**, 339–342.

Georgoulas,A. *et al*. (2012) A subsystems approach for parameter estimation of ode models of hybrid systems. In: Bartocci,E. and Bortolussi,L. (eds.) *Proceedings First International Workshop on Hybrid Systems and Biology*, EPTCS 92.

Hahn,C.N. *et al*. (2011) Heritable gata2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat. Genet.*, **43**, 1012–1017.

Hao,N. and O'Shea,E.K. (2011) Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nat. Struct. Mol. Biol.*, **19**, 31–39.

Honkela,A. *et al*. (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.

Huynh-Thu,V.A. *et al*. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Inoue,H. and Yamanaka,S. (2011) The use of induced pluripotent stem cells in drug development. *Clin. Pharmacol. Ther.*, **89**, 655–661.

Jeffreys,H. (1961) *Theory of Probability*. Clarendon Press, Oxford.

Liepe,J. *et al*. (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.*, **9**, 439–456.

Mangan,S. *et al*. (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *J. Mol. Biol.*, **356**, 1073–1081.

Moignard,V. *et al*. (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell. Biol.*, **15**, 363–372.

Moignard,V. *et al*. (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, **33**, 269–276.

Nadler,B. *et al*. (2005) Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In: Weiss,Y. *et al*. (eds.) *Advances in Neural Information Processing Systems*, **Vol. 18**.

Oates,C.J. *et al*. (2014) Causal network inference using biochemical kinetics. *Bioinformatics*, **30**, i446–i474.

Ocone,A. and Sanguinetti,G. (2011) Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, **27**, 2873–2879.

Ocone,A. *et al*. (2013) Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. *Bioinformatics*, **29**, 910–916.

O'Hagan,A. (2006) Bayesian analysis of computer code outputs: a tutorial. *Reliab. Eng. Syst. Safe.*, **91**, 1290–1300.

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

Ptashne,M. and Gann,A. (2002) *Genes and Signals*. Cold Harbor Spring Laboratory Press, New York.

Sanguinetti,G. *et al*. (2009) Switching regulatory models of cellular stress response. *Bioinformatics*, **25**, 1280–1286.

Silk,D. *et al*. (2014) Model selection in systems biology depends on experimental design. *PLoS Comput. Biol.*, **10**, e1003650.

Stathopoulos,V. and Girolami,M.A. (2012) Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philos. Trans. A Math. Phys. Eng. Sci.*, **371**, 20110541.

Trapnell,C. *et al*. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

Wang,L. *et al*. (2009) Bistable switches control memory and plasticity in cellular differentiation. *Proc. Natl Acad. Sci. USA*, **106**, 6638–6643.

Wilkinson,D.J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.*, **10**, 122–133.