Bioinformatics, 31(11), 2015, 1816–1823
doi: 10.1093/bioinformatics/btv040
Advance Access Publication Date: 31 January 2015
Original Paper

Bioimage informatics

Factor graph analysis of live cell–imaging data reveals mechanisms of cell fate decisions

Theresa Niederberger[1,†], Henrik Failmezger[2,3,†], Diana Uskat[1,†], Don Poron[3], Ingmar Glauche[4], Nico Scherf[4,5], Ingo Roeder[4], Timm Schroeder[6] and Achim Tresch[1,2,3,*]

[1]Gene Center, Department of Chemistry and Biochemistry, Ludwig-Maximilians-University München, Germany, [2]Max-Planck-Institute for Plant Breeding Research, Cologne, Germany [3]Department of Biology, Albertus-Magnus University, Cologne, Germany, [4]Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus, TU Dresden, Germany, [5]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany and [6]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Robert F. Murphy

Received on October 20, 2014; revised on December 31, 2014; accepted on January 19, 2015

Abstract

Motivation: Cell fate decisions have a strong stochastic component. The identification of the underlying mechanisms therefore requires a rigorous statistical analysis of large ensembles of single cells that were tracked and phenotyped over time.

Results: We introduce a probabilistic framework for testing elementary hypotheses on dynamic cell behavior using time-lapse cell-imaging data. Factor graphs, probabilistic graphical models, are used to properly account for cell lineage and cell phenotype information. Our model is applied to time-lapse movies of murine granulocyte-macrophage progenitor (GMP) cells. It decides between competing hypotheses on the mechanisms of their differentiation. Our results theoretically substantiate previous experimental observations that lineage instruction, not selection is the cause for the differentiation of GMP cells into mature monocytes or neutrophil granulocytes.

Availability and implementation: The Matlab source code is available at http://treschgroup.de/Genealogies.html

Contact: failmezger@mpipz.mpg.de

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

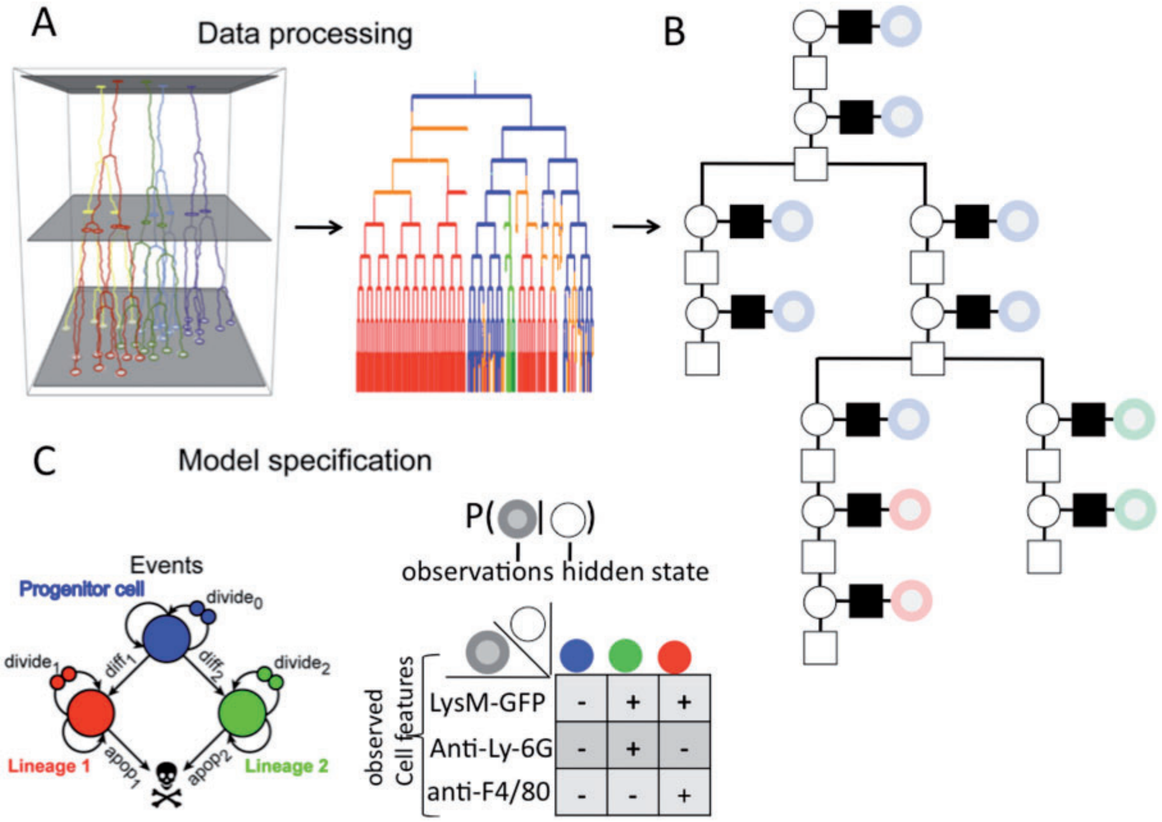
1 Introduction

Cell fate decisions appear as robust phenomena on the tissue level. However, the underlying mechanisms are hidden behind a large variability on the level of individual cells. The deterministic and the stochastic components of these decision processes can only be identified through the statistical analysis of a large number of cells. High content live cell time-lapse imaging has become a major technique for the investigation of cell behavior (Conrad and Gerlich, 2010; Neumann *et al.*, 2006, 2010; Schmid *et al.*, 2013; Starkuviene and Pepperkok, 2007). The amount of data produced by this method requires automation at all stages, starting from cell identification and tracking of cells, the extraction of relevant morphological features from single cell images, up to the biological interpretation of the results. Excellent bioinformatics tools have been developed for the purpose of cell identification, cell tracking and feature extraction (Buggenthin *et al.*, 2013; Carpenter *et al.*, 2006; Conrad *et al.*, 2011; Pau *et al.*, 2010; Rajaram *et al.*, 2012; Scherf *et al.*, 2012). So far, biological analysis of time-lapse movies focused on the detection of abnormal cell phenotypes after introducing perturbations like RNA interference (Failmezger *et al.*, 2013a; Fuchs *et al.*, 2010).

This led to the development of hidden Markov models that, either in a supervised or in an unsupervised fashion, classify cells according to their morphology and their environmental context (Failmezger *et al.*, 2013b; Snijder *et al.*, 2009; Zhong *et al.*, 2012). In this work, we develop a factor graph model, which provides a probabilistic framework for testing the compatibility of biological hypotheses with time-lapse imaging data.

Such data allow the tracking of individual cells, comprising all of their progeny over extended time periods. For each cell image, features like morphology, cell-cycle time, or motility can be extracted. All these different pieces of information on cellular development, divisional history and differentiation can be summarized into a pedigree-like structure, termed cellular genealogy, in which the founder cell represents the root, and the progeny are arranged in the branches (Fig. 1A). To our knowledge, only two attempts have been made to analyze such tree-structured data in a rigorous, probabilistic way (Durand *et al.*, 2001, 2005). Using genealogies as input data, we develop a factor graph model (Kschischang *et al.*, 2001), which describes the genealogy by a random biological process with meaningful and interpretable parameters. Factor graph models have been tremendously successful in coding theory (Chung *et al.*, 2001; Tanner, 2006) because they give rise to highly efficient algorithms for maximum likelihood estimation (MLE) and the calculation of marginal probabilities (Kschischang *et al.*, 2001). The time complexity of these algorithms is linear in the number of nodes, given that the number of neighbors of a node is bounded (by three in our case). This allows us to run a Markov Chain Monte Carlo (MCMC) algorithm for parameter learning. Our factor graph model for genealogies requires only seven parameters. Simulations show that these parameters can be identified accurately. To further assess the robustness of our model, we analyze genealogies generated by a mathematically different and considerably more complex model of stem cell differentiation (Loeffler and Roeder, 2002; Roeder and Loeffler, 2002). We are able to recover its relevant kinetic characteristics.

In our application, the factor graph model decides between alternative hypotheses concerning the mechanisms of hematopoietic progenitor cell (HPC) differentiation. It is under debate (Endele *et al.*, 2014; Glauche *et al.*, 2009; Morrison *et al.*, 1997; Rieger *et al.*, 2009; Sarrazin and Sieweke, 2011) whether cytokines instruct the differentiation of progenitor cells into specific cell types, or whether they select lineage committed cells types by allowing their proliferation or survival. Looking merely at the relative cell abundances at the end of the differentiation process, all scenarios lead to



	🔵	🟢	🔴
LysM-GFP	-	+	+
Anti-Ly-6G	-	+	-
anti-F4/80	-	-	+

Fig. 1. Conversion of a time lapse video into a factor graph. **(A)** Image processing. Left: Cell entities are identified on a series of consecutive pictures of the same area on a cell culture plate. They are linked to their respective predecessor on the previous image. Right: The topological information on cell fates is represented as a genealogy, a forest of rooted binary trees. Nodes represent cells at a certain time point, and edges indicate the parent-offspring relation from top to bottom. The vertex color informally indicates phenotypic information assigned to each cell, such as the presence or absence of fluorescent cell type markers. **(B)** Construction of the factor graph. Each node of the genealogy is represented by a hidden variable node (empty circle), and the probability of an event linking cells between consecutive time points is represented by a hidden factor node (empty square). The cell image data are represented by observable variable nodes (shaded circles). An observable factor node (black square) links each observable variable node to its corresponding hidden variable node, and it encodes the probability of observing the image data, given the hidden cell state. The factor graph encodes the model's likelihood function, the product of all factor nodes. **(C)** Specification of the local probability functions assigned to the factor nodes. We assume that each cell can be in one of three states (blue:progenitor cell state, red/green: differentiated states). Left: Each edge in the graph represents one of four events that can occur: persistence of the cell, cell division, differentiation, and apoptosis. The edges are labeled with the (unknown) probabilities for the respective event, which determine the probability functions of the hidden factor nodes. Right: The functions related to the observable factor nodes are conditional probability distributions of the single cell image data, given the cell's state

indistinguishable outcomes, namely the dominance of one cell type. We construct a reversible-jump MCMC algorithm, which is able to identify the most realistic scenario as well as the corresponding lineage specific differentiation, proliferation and cell death rates. Our analysis of a HPC time-lapse imaging experiment (Rieger *et al.*, 2009) identifies unequal proliferation as the driving force for the asymmetric differentiation of murine granulocyte-macrophage progenitors (GMPs) cells into either mature monocytes or neutrophil granulocytes.

2 Methods

2.1 Data acquistion

The data in Rieger *et al.* (2009) consist of genealogies generated from live cell imaging of murine GMPs cells in the presence of only macrophage- or granulocyte colony-stimulating factor. Cells contained a LysM::GFP marker, expressing enhanced green fluorescent protein (GFP) from the lysozymeM gene locus as an early molecular reporter for unilineage commitment. Weak lysozymeM::GFP expression is found in undifferentiated GMPs, whereas its expression is drastically up-regulated in differentiated cells. At the end of the experiment, stainings were performed according to standard immunofluorescence procedures (see Methods in Rieger *et al.*, 2009). Cell morphology and Anti-Ly-6G (Miltenyi Biotech) and anti-F4/80 (eBioscience) were used for the identification of granulocytes and macrophages, respectively. Absence or presence of a marker was assessed manually by an expert. Potential discretization errors when converting a continuous fluorescence signal to a binary signal were accounted for in Equation (5) of the main text. To avoid biases from incomplete tracking, cell division events in which only one daughter cell could be tracked were discarded by cutting the tree above the event.

2.2 Modeling of genealogies as factor graphs

Time-lapse movies encode two principally different types of information. Apart from single-cell image data at each time point, movies track cellular genealogies, i.e. they record the history of a cell population at the individual cell level. Let V be the set of all single cell images that have been acquired. Image analysis yields a set of statistical features o_v, $o = \{o_v | v \in V\}$, derived from each single cell image. Additionally, through cell tracking, we obtain a collection Γ of rooted binary trees with node set V. An edge $v \to w$ is drawn whenever v and w are cell images in consecutive frames of the movie, and if w shows either the same cell as v or an offspring of v (see Fig. 1A). For an edge $v \to w$, v is called a parent of w, and w is called a child or offspring of v. The (possibly empty) set of children of a node v is denoted by $ch(v)$.

Our key concept is to model dynamic cellular processes as probabilistic transitions between discrete 'states' S of a cell. Because the cell states typically cannot be observed directly, we call these states hidden. For HPC differentiation, e.g. we assume that each cell at a given time point is either in a undifferentiated (blue) cell state, or it is in one of two (red respectively green) differentiated cell states (Fig. 1A). Hypotheses on the differentiation mechanism can be easily formulated in terms of these states: do red cells die faster than green cells? Do blue cells preferentially develop into green cells rather than red cells? In each time interval between two images, i.e. along each edge of the genealogy, one of the following events can occur: most likely, a cell will persist in its current state. Alternatively, the cell may divide, die, or, in case of a progenitor cell, differentiate into a red or a green cell. Each of these events has its own, unknown probability (Fig. 1C). We have only indirect information on a cell's state, given by the features extracted from its cell image, such as the fluorescence intensity of a progenitor cell marker. Those features provide evidence for or against a certain cell state, since each cell state has a characteristic feature distribution. The learning of the unknown event probabilities from uncertain information is a standard task in statistical learning. If the genealogies were linear, i.e. if the cells never divided, the classical approach to our problem would be a hidden Markov model (Held *et al.*, 2010). However, cell fate and cellular decision making are intimately linked to cell division events, for which reason we need to model non-linear genealogies. Our factor graph model can be viewed as a generalization of the hidden Markov model to network topologies. Noteworthy, our factor graph model does not belong to the class of Bayesian networks, as Bayesian networks require cell states of two daughter cells to be conditionally independent given their parent cell's state. This property is often violated in practice, asymmetric cell division in *Saccharomyces cerevisiae* being a well-known example (Lord and Wheals, 1980).

The factor graph assigned to a genealogy consists of two node types, variable nodes and factor nodes (Fig. 1B). To each cell image $v \in V$, we define a variable node H_v representing the cell's hidden state. A parent cell v and its daughter cell(s) $ch(v)$ are linked by a factor node f_v which encodes a probability function of its adjacent variable nodes (Fig.1C). Here, $f_v(H_{ch(v)}, H_v; \theta)$ is the probability that a cell v of type $H_v \in S = \{\text{blue}, \text{green}, \text{red}\}$ will give rise to 0, 1 or 2 offspring of type $H_{ch(v)}$ in the next time step. Here, θ denotes the parameters of our model. Motivated by the fact that the cell states S cannot be observed directly, we call this part of the model the hidden layer, and its nodes are called the hidden nodes. The second, observable layer consists of (observable) variable nodes O_v, $v \in V$, which represent the image features extracted from v. The observable node O_v is linked to its corresponding hidden node H_v by a factor node g_v encoding a probability function of its adjacent variable nodes (Fig. 1C). Here, the emission function $g_v(O_v, H_v; \theta)$ encodes the probability of observing the image features O_v, given that the hidden state of cell v is H_v. The graph we have constructed is bipartite in the sense that factor nodes are connected to variable nodes only, and vice versa. Let $H = (H_v)_{v \in V}$, $\mathcal{O} = (O_v)_{v \in V}$. We assume that the joint probability of $\mathcal{O}$, H and Γ decomposes according to the factor graph topology, i.e. it is the product of its factor nodes g_v and h_v:

$$\begin{aligned} P(\mathcal{O}, H; \theta) &= P(\mathcal{O} | H; \theta) \cdot P(H; \theta) \\ &= \prod_{v \in V} \underbrace{P(O_v | H_v; \theta)}_{=: g_v(O_v, H_v; \theta)} \cdot \prod_{v \in V} \underbrace{P(H_w, w \in ch(v) | H_v; \theta)}_{=: f_v(H_{ch(v)}, H_v; \theta)} \end{aligned} \quad (1)$$

We point out that the factor graph representation of the joint probability in Equation (1) does not agree with the factorization induced by the interpretation of the genealogy as a Bayesian network and its subsequent canonical conversion into a factor graph (see Kschischang *et al.*, 2001, Section B)

2.3 Parametrization of the factor graph model

First, we need to choose the number $|S|$ of cell states, one for each cell type, and then define the possible events (cell division, differentiation, apoptosis, no event) that may occur between two observation time points (Fig. 1B). The probabilities for these events are encoded in the functions $f_v(H_{ch(v)}, H_v; \theta) = P(H_w, w \in ch(v) | H_v; \theta)$. Formally, this requires $|S|$, $|S| \cdot (|S| - 1)$, and $|S| \cdot (|S|^2 - 1)$ parameters for respectively $|ch(v)| = 0, 1, 2$ daughter cells of a cell v. In our application, S has three elements, which amounts to 32 parameters. However, prior knowledge can reduce this number substantially. The probability of an apoptosis event ($ch(v) = \emptyset$) is given by

$$f_v(H_v = s;\, \theta) = \mathrm{apop}_s\,,\ s \in S \tag{2}$$

We set $\mathrm{apop}_{\mathrm{blue}} = 0$, because we do not observe progenitor cell death events during the observation period. If f_v represents a cell division event ($|ch(v)| = 2$),

$$f_v(H_{ch(v)} = (s_1, s_2),\, H_v = s;\, \theta) = \begin{cases} \mathrm{divide}_s & \text{if } s_1 = s_2 = s \\ 0 & \text{else} \end{cases} \tag{3}$$

In Equation (3), we are assuming that the daughter and parent cells in cell division events are in the same state. If f_v represents persistence or differentiation ($|ch(v)| = 1$), we let

$$f_v(H_{ch(v)} = t,\, H_v = s;\, \theta)$$
$$= \begin{cases} \mathrm{diff}_t & \text{if } s = \mathrm{blue}, t \neq \mathrm{blue} \\ 1 - \mathrm{diff}_{\mathrm{red}} - \mathrm{diff}_{\mathrm{green}}\, _{- \mathrm{divide}_{\mathrm{blue}} - \mathrm{apop}_{\mathrm{blue}}} & \text{if } s = \mathrm{blue}, t = \mathrm{blue} \\ 1 - \mathrm{divide}_s - \mathrm{apop}_s & \text{if } s = t \neq \mathrm{blue} \\ 0 & \text{else} \end{cases} \tag{4}$$

The latter choice is motivated by the fact that only progenitor cells (blue state) are able to change their state into one of the differentiated cells. Altogether, the hidden layer of our factor graph model is determined by only seven parameters, $\theta = \{\mathrm{divide}_{\mathrm{blue}}, \mathrm{divide}_t, \mathrm{apop}_t, \mathrm{diff}_t;\ t = \mathrm{red},\ \mathrm{green}\}$.

Second, we have to relate the cell states to the image data. Several techniques have been developed to identify these states either by *in vivo* fluorescence staining or by immunostaining after fixation. All these methods have their drawbacks: the marker might not work with equal efficiency in all cells. Fluorescence intensity is a continuous marker, and converting it to a discrete signal can cause discretization errors. The information provided by these markers may be incomplete, e.g. in the case of the Lysm marker in our experimental data (Rieger *et al.*, 2009), the upregulation of Lysm fluorescence signal merely indicates that the cell is no longer in the progenitor cell state, but it does not tell whether it has become a red or a green cell. The second marker (F4/80) is detectable only if the cell is a monocyte, but it does not distinguish between progenitor cells and granulocytes. Using the combination of both markers, we have a ternary yet error-prone readout indicating the state of each cell. We model this by letting

$$g_v(o_v, H_v;\, \theta) = P(o_v | H_v;\, \theta)$$

$$P(o_v | H_v;\, \theta) = \begin{cases} 0.8 & \text{if readout } o_v \text{ is indicative of state } s, \text{ and } H_v = s \\ 0.1 & \text{if readout } o_v \text{ not indicative of state } s, \text{ and } H_v \neq \mathrm{s} \end{cases} \tag{5}$$

We verified that the concrete choice of the numerical values in Equation (5) is of minor importance and does not influence our final model decision qualitatively (data not shown).

2.4 Parameter estimation and hypothesis testing

It is crucial that the factor nodes encode ‘local’ probability functions in the sense that they depend only on their respective neighboring variable nodes. Because the hidden states are unknown, we are interested in calculating the marginal likelihood,

$$L(\theta) = P(\mathcal{O};\, \theta) = \sum_h P(\mathcal{O}, H = h;\, \theta) \tag{6}$$

$L(\theta)$ is obtained by summation of the full likelihood over all possible hidden state combinations $h = (h_v)_{v \in V}$, $h_v \in S$. Factor graphs give rise to efficient algorithms for MLE (max-sum algorithm) and the calculation of marginal probability distributions (sum-product algorithm) which are linear in the number of nodes see, e.g. (Bishop, 2006, Chapter 8) as long as the number of neighbors of a factor nodes is bounded. The processing of genealogies with thousands of variables with the standard algorithms is numerically unstable. The calculations were therefore implemented in log space (see Niederberger *et al.*, 2012) Supplementary Material Section S2). There are two main strategies for parameter estimation: Point estimation methods, foremost MLE, and sampling methods like MCMC sampling. Point estimation methods typically are fast, yet they suffer from the danger of getting trapped in a local maximum of the likelihood function. Moreover, it is not easy to construct an MLE estimator for factor graph models (e.g. using Expectation-Maximization (Dempster *et al.*, 1977). We therefore implemented a Metropolis-Hastings MCMC approach (Supplemental Methods S2.1 and S2.2), which generates a sequence $\Theta = (\theta, \theta_2, ..., \theta_T)$ of parameter values drawn according to the likelihood function $L(\theta)$. For T large enough, the empirical distribution of Θ is representative of $L(\theta)$, i.e. it converges in the weak sense towards the distribution defined by $L(\theta)$. In our applications, we chose $T = 20$ 000, and trace plots were used to verify the convergence of the Markov chain (Fig. 2C).

Although parameter estimates alone can already provide useful information, the strength of our model lies in the opportunity to formulate biological hypothesis in terms of these parameters. A biological hypothesis $\mathcal{H}$ can be compatible or incompatible with the parameter sample θ. In our application, we address the question whether asymmetric differentiation of progenitor cells (blue state) into two different mature cell types (green and red state) is achieved by selection or instruction. In a hypothetical selective scenario $\mathcal{H}_{\mathrm{sel}}$, progenitor cells divide into red and green state cells with roughly the same probability, but the apoptosis rate of green and red state cells may differ. As an alternative hypothesis, in an instructive scenario $\mathcal{H}_{\mathrm{inst}}$, apoptosis rates of the mature cells are approximately equal, but the probability for progenitor cells to differentiate into a green or red state cell may differ. Both hypotheses can be formulated in terms of the factor graph model parameters. In the selective scenario $\mathcal{H}_{\mathrm{sel}}$, we assume $\mathrm{divide}_{red} = \mathrm{divide}_{green}$, whereas in the instructive scenario $\mathcal{H}_{\mathrm{inst}}$, we assume and $\mathrm{apop}_{red} = \mathrm{apop}_{green}$. Thus, the space Θ_{inst} of parameters compatible with $\mathcal{H}_{\mathrm{inst}}$ is different from the parameter space Θ_{sel} compatible with $\mathcal{H}_{\mathrm{sel}}$. We use a reversible jump MCMC algorithm for the sampling $\theta_1, \theta_2, ...$ from $\Theta_{\mathrm{inst}} \cup \Theta_{\mathrm{sel}}$ (see Supplementary Material S2.3), which can switch between the two parametrizations Θ_{inst} respectively Θ_{sel} of the factor graph model.The decision between two competing hypotheses is based on the ratio $\log\left(\frac{|\{\theta_j | \theta_j \in \Theta_{\mathrm{inst}}\}|}{|\{\theta_j | \theta_j \in \Theta_{\mathrm{sel}}\}|}\right)$. Large values provide evidence for the instructive, small values provide evidence for the selective scenario. The code for the sum-product algorithm and for the reversible jump MCMC algorithm Supplemental Methods S2.3.

3 Results

3.1 Parameters can be identified from a moderate number of genealogies

Given the experimental data $(\mathcal{O}, \Gamma)$, our goal is to identify the model parameters (the unknown probabilities characterizing cell-type specific behavior) with sufficient precision to draw biological conclusions from it. The success of the method depends on its discriminatory power and on the amount of available data. We have addressed both issues in an extensive simulation study. We check whether a ‘true’ set of parameters can be recovered from data, which

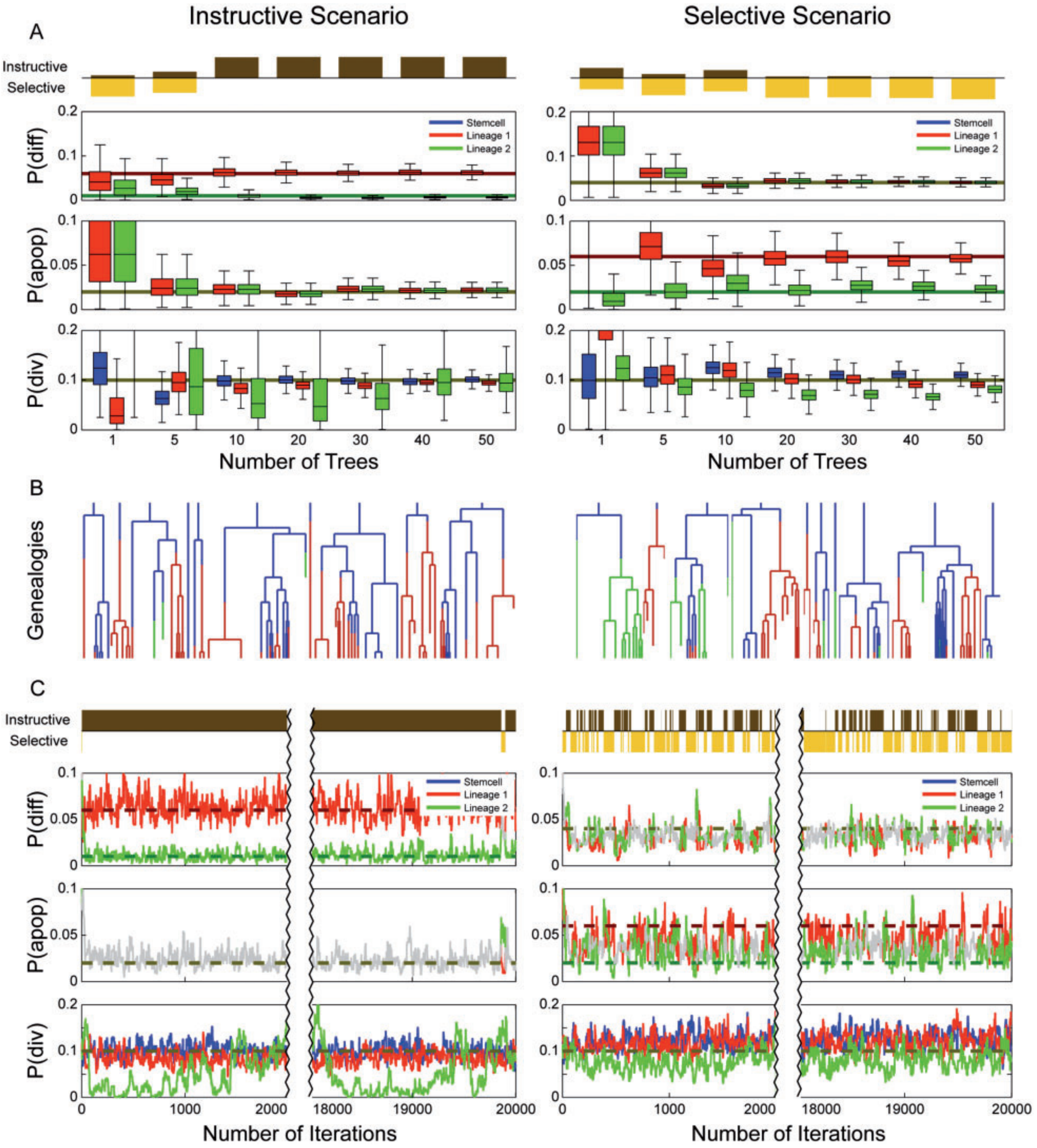


Fig. 2. Parameter sampling and hypothesis discrimination by reversible jump MCMC in a simulation of an instructive scenario (left-hand side) and a selective scenario (right-hand side). **(A)** For each scenario, we simulated experiments with different numbers of trees (1, 5, 10, 20, 30, 40, 50). Boxplots of the predicted differentiation, apoptosis and division probabilities, as well as the proportions of the predicted scenario depending on the number of trees are shown. The horizontal lines depict the simulated ('true') parameter values which were used for the generation of these simulated genealogies. On the x-axis the numbers of trees used for the predictions are depicted in increasing order whereas each dataset is a subset of the next one in size. The barplot at the top pictures the scenario proportion, where instructive is colored in brown and the selective in yellow. The boxplots are divided into three subgroups according to the predicted probabilities for differentation (top), apoptosis (middle) and division (bottom) whereas the used colors correspond to the three different lineages. Note that for |trees| = 1 the boxplot boxes may exceed the displayed range of values. **(B)** Simulated trees for selective scenario and instructive scenario, which are used for the prediction (here: |trees| = 10). The colors correspond to the one in (A) where progenitor cells are blue and the two different lineages are red and green. **(C)** Trace plots for all predicted parameters and the jumps between the two scenarios (here:|trees| = 10), to ensure convergence of the Markov chain. Note that only the beginning and the end of the 20 000 MCMC iterations are shown. The barplot (top) describes the reversible jump between selective and instructive by depicting the respective scenario for each given iteration step. The colors for the jump representation and the trace plots as well as for the parameters correspond to those in (A). Grey indicates that the corresponding parameter is the same for both lineages in the currently selected model class. The horizontal dashed lines visualize the simulated ('true') parameter values

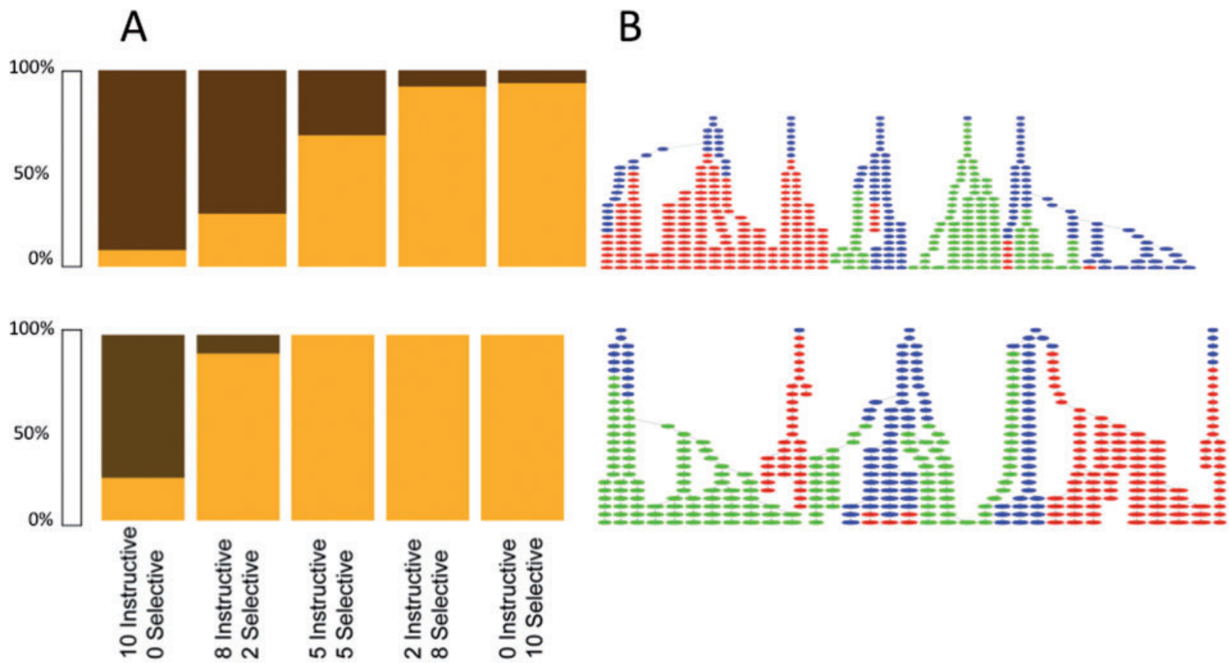


Fig. 3. Hypothesis discrimination between selective and instructive scenarios. Discrimination on simulated data from the factor graph model (upper row) and from the model in (Loeffler and Roeder, 2002; Roeder and Loeffler, 2002) (lower row) **(A)**. 10 trees were used in each of 100 MCMC runs. The 10 trees consisted of $j = 10, 8, 5, 2, 0$ (left column to right column) trees taken from a selective scenario, and $10 - j$ trees from a selective scenario. The illustration shows the percentage of cases in which the MCMC algorithm preferred the selective (yellow) or the instructive scenario (brown). Examples of simulated genealogy trees for the instructive scenarion (upper row, $j = 10$) and the selective scenario (lower row, $j = 5$) are shown in **(B)**

has been simulated using these parameters. We generated data sets of different sizes and ran the MCMC estimation procedure, which gave us credible intervals for the individual parameters. As expected, the accuracy of parameter estimation increased with the amount of data. It turned out that for realistic parameter choices (see Fig. 2 and Supplementary Figs. S2 and S3), a moderate number of about 20 genealogies was sufficient for accurate parameter identification and a reliable decision between the instructive and the selective scenario. This is substantially less than the ~200 genealogieswith an average of 12 cell division events in the experimental data.

The simulation also allowed us to exclude severe estimation biases resulting from our MCMC procedure. By mixing genealogies from a selective and an instructive scenario at different ratios, we verify that the estimation procedure is robust. It generally tends to decide for the scenario from which the majority of the genealogies were taken from (see Fig. 3). The results also revealed that the estimation procedure is slightly biased towards the selective scenario. This has to be taken into account when evaluating the results from experimental data.

Hypothesis discrimination in realistic models of cell differentiation

So far, we merely verified the reliability of our method in the absence model bias, i.e. if the data used for parameter learning was generated from the factor graph model itself. However, our model is a gross simplification of the real biological processes. We therefore tested the performance, in particular the model bias, in a more realistic model of hematopoietic stem cell organization (Loeffler and Roeder, 2002; Roeder and Loeffler, 2002; Glauche *et al.*, 2007). In this model, stem cell differentiation is described as a temporally extended process with a progressive restriction in lineage potential, in which the decision process is intrinsically random, albeit tuneable by lineage instruction or selection. Applying our analysis to a set of model-generated genealogies, we observe that the true scenario (selective/instructive) can be recovered from merely 10 trees in ~90% of all cases. Again, we mixed genealogies generated by the two scenarios at different ratios and monitored the decision of our method. The results confirmed a slight bias towards the selective scenario (Supplementary Fig. S3). The most likely explanation is that selective scenarios can be recognized by a combination of fluorescence signal and topological information (shorter branches in the tree are indicative of apoptosis events), while evidence for instructive scenarios arises merely from the fluorescence signal. Because the Loeffler/Roeder/Glauche model predicts uneven apoptosis probabilities for the two differentiated cell types, even in an instructive scenario, this leads to the observed bias. We emphasize that our biological conclusions in the subsequent application are unaffected, because they point in opposite direction, towards an instructive scenario.

Preferential differentiation of murine GMPs into monocytes or neutrophil granulocytes cells is achieved by an instructive mechanism

It is still under debate how the commitment of multipotent HPCs to single lineages is controlled (Endele *et al.*, 2014; Morrison *et al.*, 1997; Glauche *et al.*, 2009; Rieger *et al.*, 2009). Cytokines are known to influence cell fates, but the driving mechanism is still unclear: Is lineage commitment achieved only in a selective manner, by allowing the survival and proliferation of cells belonging to one lineage, or are they able to directly instruct lineage commitment? (Rieger and Schroeder 2009) tracked the development of murine GMPs (blue state) into mature monocytes (M, red state) or neutrophil granulocytes (G, green state) on the single-cell level (Rieger *et al.*, 2009). We applied our factor graph model to genealogies

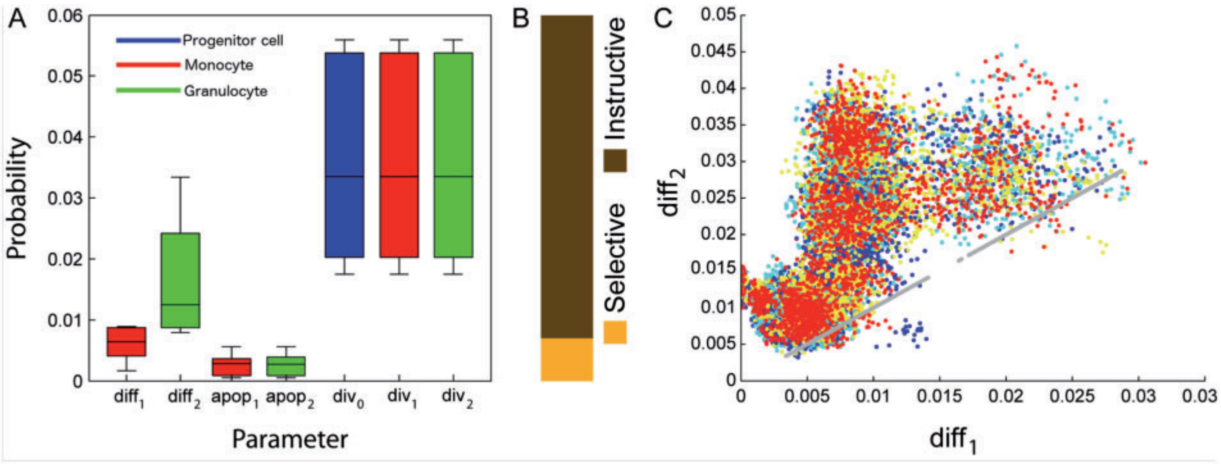


Fig. 4. Parameters estimated from the progenitor cell differentiation data **(A)**. Boxplots summarize the mean parameter values obtained from 10 independent MCMC runs. **(B)** The instructive scenario was preferred in 88% of all steps over the selective scenario. **(C)** Plot of the joint parameter values ($\text{diff}_{\text{rednbsp;}}$, $\text{diff}_{\text{green}}$) for 4 MCMC runs (yellow, light/dark blue, red). Grey dots on the main diagonal indicate steps in which the MCMC chain was in a selective scenario

from this data set to shed light onto the driving mechanism (see e.g. Supplementary Fig. S4). Our genealogy consists of 200 trees constructed from data as provided by the authors of the study (Rieger *et al.*, 2009). This amount of data could not be processed by our method as a whole. The data were randomly split into 10 disjoint data sets with 20 trees each. Consistent with the findings obtained by visual inspection in (Rieger *et al.*, 2009), the factor graph model identifies an instructive mechanism as the cause of a strong preferential differentiation into monocytes or granulocytes, respectively. The reversible jump MCMC favors the instructive scenario over the selective scenario in 88% (Fig. 4B). Recall that our model has a slight estimation bias towards selective scenarios, so we can exclude bias as the cause for our findings. In the instructive scenario, the probability for the differentiation of a progenitor cell into a granulocyte is estimated as 0.012 per time step (median of all sampled probabilities) versus 0.007 per time step for monocytes (Fig. 4A). This difference becomes even more evident in Figure 4C, which shows a scatterplot of the two differentiation probabilities along four MCMC chains. The sampled values of diff2 exceed the corresponding values of diff1 in all but very few exceptions. The probability of going into apoptosis is estimated as 0.002 (median of sampled probabilities) and thus substantially smaller than the differentiation probabilities. Further, the cell type specific growth speed, i.e. the division rate was remarkably similar (Fig. 4A).

4 Conclusion

High throughput time-lapse imaging poses new challenges to computational biology. The analysis of cellular genealogies requires the development of statistical methods that are able to test hypotheses on data with a branched dependency structure. The most advanced method so far that deal with this kind of data is Markov tree models (Durand *et al.*, 2001, 2005). They can be cast as factor graph models, in which the fate of the children is assumed independent of each other, given the (state of the) parent cell. This assumption has the advantage that Markov trees allow the straightforward generalization of the Baum-Welch algorithm for parameter learning. However in cases when diverging daughter cell fates are coupled, this model is not appropriate. We have developed a simplistic, but efficient factor graph model, which can test the compatibility of basic biological hypotheses with this kind of data. The factor graph model has an intuitive parametrization of its factor nodes, which reflects key properties of the cellular system (differentiation, proliferation, apoptosis). Biological hypotheses can be formulated in terms of parameter restrictions. The local factor nodes and their parametrization can be exchanged very easily in this framework. Thus, our model can be easily adapted to model high-dimensional readouts from each cell image, or to model other cell states (aneuploidy, cancerosity, quiescence). We perform hypothesis discrimination and parameter estimation using a reversible jump MCMC algorithm. More efficient algorithms for parameter estimation such as an EM-algorithm would be convenient; they are currently under development. Due to its flexibility, we expect the factor graph framework to have a wide range of applications, in particular because advances in microscopy allow the investigation of cell-to-cell (state) variability at the subcellular level.

The factor graph model can be extended to account for the micro-environment of a cell, given by cell contacts, cell wall turgor, local concentration of signaling molecules, or concentration gradients. Such variables might be included into the observations vector o_ν for each cell image ν. A promising application will be the modeling of multicellular organism development. For example, plant embryos can be tracked, and at the same time chemo-physical properties of each cell can be measured (Kierzkowski *et al.*, 2012), which can reveal to which extent lineage, mechanical forces or location within a tissue determine cell fate.

Acknowledgement

We thank Carsten Marr for help with the processing of the data published in Rieger *et al.*, 2009).

Funding

T.N. and A.T. were supported by an SFB 646 grant from the Deutsche Forschungsgemeinschaft.

Conflict of Interest: none declared.

References

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning. Information Science and Statistics*. New York, Springer.

Buggenthin,F. *et al.* (2013) An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy. *BMC Bioinformatics*, **14**, 297.

Carpenter,A.E. *et al.* (2006) Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, **7**, R100.

Chung,S.-Y. *et al.* (2001) On the design of low-density parity-check codes within 0.0045 db of the shannon limit. *Commun. Lett. IEEE*, **5**, 58–60.

Conrad,C. and Gerlich,D.W. (2010) Automated microscopy for high-content rnai screening. *J. Cell Biol.*, **188**, 453–461.

Conrad,C. *et al.* (2011) Micropilot: automation of fluorescence microscopy-based imaging for systems biology. *Nat. Methods*, **8**, 246–249.

Dempster,A.P. *et al.* (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. B.*, **39**, 1–38.

Durand,J.-B. Goncalves,P. (2001) Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees. [Research Report] RR-4248, 2001. <inria-00072339>.

Durand,J.-B. *et al.* (2005) Analysis of the plant architecture via tree-structured statistical models: the hidden markov tree models. *New Phytologist*, **166**, 813–825.

Endele,M. *et al.* (2014) Instruction of hematopoietic lineage choice by cytokine signaling. *Exp. Cell Res.*, **329,** 207–213.

Failmezger,H. *et al.* (2013a) Learning gene network structure from time laps cell imaging in rnai knock downs. *Bioinformatics*, **29**, 1534–1540.

Failmezger,H. *et al.* (2013b) Unsupervised automated high throughput phenotyping of rnai time-lapse movies. *BMC Bioinformatics*, **14**, 292.

Fuchs,F. *et al.* (2010) Clustering phenotype populations by genome-wide rnai and multiparametric imaging. *Mol. Syst. Biol.*, **6**, 370.

Glauche,I. *et al.* (2007) Lineage specification of hematopoietic stem cells: mathematical modeling and biological implications. *Stem Cells*, **25**, 1791–1799.

Glauche,I. *et al.* (2009) A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell Prolif.*, **42**, 248–263.

Held,M. *et al.* (2010) Cellcognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. Methods*, **7**, 747–754.

Kierzkowski, D. *et al.* (2012) Elastic domains regulate growth and organogenesis in the plant shoot apical meristem. *Science*, **335**, 1096–1099.

Kschischang,F.R. *et al.* (2001) Factor graphs and the sum-product algorithm. Inform. Theory, IEEE Trans., **47**, 498–519.

Loeffler,M. and Roeder,I. (2002) Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models–a conceptual approach. *Cells Tissues Organs*, **171**, 8–26.

Lord,P.G. and Wheals,A.E. (1980) Asymmetrical division of saccharomyces cerevisiae. *J. Bacteriol.*, **142**, 808–818.

Morrison,S.J. *et al.* (1997) Regulatory mechanisms in stem cell biology. *Cell*, **88**, 287–298.

Neumann,B. *et al.* (2006) High-throughput rnai screening by time-lapse imaging of live human cells. *Nat. Methods*, **3**, 385–390.

Neumann,B. *et al.* (2010) Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, **464**, 721–727.

Niederberger,T. *et al.* (2012) Mc eminem maps the interaction landscape of the mediator. *PLoS Comput. Biol.*, **8**, e1002568.

Pau,G. *et al.* (2010) Ebimage-an r package for image processing with applications to cellular phenotypes. *Bioinformatics*, **26**, 979–981.

Rajaram,S. *et al.* (2012) Phenoripper: software for rapidly profiling microscopy images. *Nat. Methods*, **9**, 635–637.

Rieger,M.A. and Schroeder,T. (2009) Instruction of lineage choice by hematopoietic cytokines. *Cell Cycle*, **8**, 4019–4020.

Rieger,M.A. *et al.* (2009) Hematopoietic cytokines can instruct lineage choice. *Science*, **325**, 217–218.

Roeder,I. and Loeffler,M. (2002) A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. *Exp. Hematol.*, **30**, 853–861.

Sarrazin,S. and Sieweke,M. (2011) Integration of cytokine and transcription factor signals in hematopoietic stem cell commitment. Semin. Immunol., **23**, 326–334.

Scherf,N. *et al.* (2012) Imaging, quantification and visualization of spatio-temporal patterning in mesc colonies under different culture conditions. *Bioinformatics*, **28**, i556–i561.

Schmid,B. *et al.* (2013) High-speed panoramic light-sheet microscopy reveals global endodermal cell dynamics. *Nat. Commun.*, **4**, 2207.

Snijder,B. *et al.* (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, **461**, 520–523.

Starkuviene,V. and Pepperkok,R. (2007) The potential of high-content high-throughput microscopy in drug discovery. *Br. J. Pharmacol.*, **152**, 62–71.

Tanner,R. (2006) A recursive approach to low complexity codes. *IEEE Trans. Inf. Theor.*, **27**, 533–547.

Zhong,Q. *et al.* (2012) Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nat. Methods*, **9**, 711–713.