BEYOND GENEALOGIES: MUTUAL INFORMATION OF CAUSAL PATHS TO ANALYSE SINGLE CELL TRACKING DATA

Nico Scherf[⋆]*, Thomas Zerjatke*[⋆]*, Konstantin Klemm*[†]*, Ingmar Glauche*[⋆] *and Ingo Roeder*[⋆]

[⋆] Institute for Medical Informatics and Biometry, Dresden University of Technology, Fetscherstr. 74, D-01307 Dresden, Germany

[†] Bioinformatics Group, Institute for Computer Science, University of Leipzig, Härtelstr. 16-18, D-04107 Leipzig, Germany

ABSTRACT

Single cell tracking, based on the computerised analysis of time-lapse movies, is a sophisticated experimental technique to quantify single cell dynamics in time and space. Although the resulting *cellular genealogies* comprehensively describe the divisional history of each cell, there are many open questions regarding the statistical analysis of this type of data. In particular, it is unclear, how tracking uncertainties or spatial information of cellular development can correctly be incorporated into the analysis. Here we propose a generalised description of single cell tracking data by spatiotemporal networks that accounts for ambiguities in cell assignment as well as for spatial relations between cells. We present a way to measure correlations among cell states by analysing the mutual information in state space considering causal (time-respecting) paths and illustrate our approach by a corresponding example. We conclude that a comprehensive spatiotemporal description of single cell tracking data is ultimately necessary to fully exploit the information obtained by time-lapse imaging.

***Index Terms*—** cell tracking, lineage trees, temporal networks, information theory, stem cells

1. INTRODUCTION

Time-lapse microscopy is an extremely valuable technique for addressing basic questions in stem cell biology [1, 2], development [3] and regenerative medicine, as it naturally allows to study single cell behaviour in space and time [2]. Many different methods for partially automated cell tracking have been proposed in recent years, cf. [4]. The typical result from single cell tracking are lineage trees, also termed *cellular genealogies*, comprising the divisional history of each cell and its progeny along with annotated data on time of division, cell position and further cell specific parameters. However, this genealogy-based description has two major shortcomings: (i) Ambiguities in cell tracking between consecutive frames are not sufficiently represented as only the most likely assignment between subsequent cell representations is contained in a particular genealogy. This implies that the effect of possible tracking errors on the resulting trees and the statistics derived therefrom can hardly be addressed. (ii) Positional information of cells is not appropriately represented to allow for analysing the influence of spatial interactions (that may often play an important role [5]). In order to address these shortcomings we present a novel approach to generalise the notion of cellular genealogies. We use spatiotemporal networks to describe alternate developmental paths induced by ambiguous cell assignments. This framework can further incorporate spatial interactions among cells and can be used to study a range of questions not assessable on the basis of traditional genealogies.

This work was supported by the German Ministry for Education and Research, BMBF (BMBF-FKZ 0315452, *HaematoSys*), the German Research Foundation, DFG (RO3500/2-1), and the Human Frontier Science Program, HFSP (RGP0051/2011.

2. RELATED WORK

A number of works focused on the analysis of cell cycling dynamics based on cellular genealogies [6, 7, 8]. Furthermore the synchronicity of cycling between related cells (e.g. daughters and siblings) has been studied in [6, 9]. Nordon et al. [8] further introduced an analysis of genealogical trees based on the statistics of branching processes. Apart from that, only few works have addressed more sophisticated measures of cellular development obtained from genealogies [10, 11]. In [10] we were among the first to propose a set of different measures to reliably detect asymmetries in cellular genealogies and potential correlations between apoptotic events of related cells. In particular, the latter analysis provided an important step to study regulation of differentiation in stem cell cultures. However, to the best of our knowledge, no work has explicitly addressed the possible ambiguities in cell assignments during cell tracking and their influence on the interpretation of cellular developments.

3. METHODS

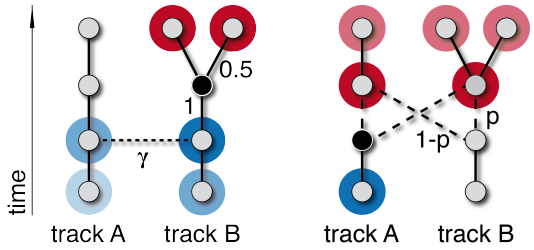
The basic rationale behind our approach is the following: Given the recorded development of a cell population over time we are interested in uncovering correlations between cells that share common features, such as lineage fate, or apoptosis. In order to address the causal relations of individual cellular developments we employ a mutual information measure. This in turn requires the definition of cellular states (e.g. fates assigned to individual cell objects) and a representation of the relations between those objects/states. These relations can either be a genealogical coupling (i.e. cells share a common ancestor) or a spatial coupling (i.e. cells had contact with each other). The resulting spatiotemporal network structures comprise both information and are used to calculate all *causal paths* between cells. Simply speaking, two cells are connected by a causal path if they share a potential source of influence. For a particular cell of interest at time point τ, the areas of potential influence can be defined similar to the concept of *light cones* (see Fig.1). Based on this concept of causal paths we can now calculate the empirical co-occurrence of two (causally connected) cells in defined states, which is the essential information to be contained in our required *state space matrix*. Based on this matrix we calculate the mutual information between any two detected cell objects. A toy example illustrating the conversion of the spatiotemporal network into the matrix representation is provided in Fig.2.

The described spatiotemporal representation is naturally suited to also incorporate the possible, technical ambiguity in cell tracking (Fig.1b). In these cases the temporal assignments of cell objects (vertical links) are replaced by the corresponding probabilities for their accuracy, thus accounting for a multitude of possible developments that cannot be excluded based on the tracking method. In this paper we will concentrate on influence of potential tracking ambiguities on the mutual information measure. The formal structure of our approach is provided below.

3.1. Spatiotemporal networks

Single cell tracking data from time lapse experiments can be represented by means of specific network structures that we denote as *spatiotemporal networks* $\mathcal{G} = \mathcal{G}(\mathcal{N}, A, C)$. We define a set of nodes $\mathcal{N}$, where each node, labelled by the tuple (i, t), represents the observation of a cellular object i at timepoint t. The development of cells over time is described as a set of edges $A = \cup A^{(t)}$ represented by the following temporal adjacency matrices (for each time point t) :

$$A^{(t)}_{i,j} = \begin{cases} 1, & \text{if } (j,t) \text{ and } (i,t-1) \text{ represent the} \\ & \text{same cell at consecutive time-points} \\ \frac{1}{2}, & \text{if } (j,t) \text{ is a daughter cell of } (i,t-1) \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$



(a) (b)

Fig. 1. **Spatiotemporal networks:** Principal architecture of spatiotemporal networks incorporating spatial contacts (a) or tracking ambiguities (b). Furthermore, examples for areas of influence (past/blue and future/red) are highlighted for the cells marked in black. Shading intensity of coloured disks correspond to temporal distance.

This network representation of genealogies can now be adapted to deal with tracking ambiguities by transforming $A^{(t)}$ into probability matrices (cf. Fig.1b):

$$\mathcal{A}^{(t)}_{i,j} = \begin{cases} p, & \text{if } (j,t) \text{ and } (i,t-1) \text{ represent the} \\ & \text{same cell with probability } p \\ \frac{p}{2}, & \text{if } (j,t) \text{ is a daughter cell of } (i,t-1) \\ & \text{with probability } p \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The temporal network can be extended to represent spatial relationships C between cells by additionally defining spatial adjacency matrices for all time points t:

$$C^{(t)}_{i,j} = \begin{cases} \gamma, & \text{if cells } i \text{ and } j \text{ interact at time } t \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

An interaction might be a direct cell-cell contact or spatial proximity with γ being the interaction strength, see Fig.1a.

The resulting spatiotemporal network $\mathcal{G} = \mathcal{G}(\mathcal{N}, A, C)$ is a comprehensive representation of single cell tracking data that contains genealogical information as well as potential spatial interactions in one framework. By replacing temporal assignments of cell objects with corresponding probabilities for their accuracy as in equation (2), one is able to reflect uncertainties of the single cell tracking. In the following we will confine ourselves to the temporal network component, but the analysis can also incorporate the spatial component.

3.2. Mutual information of state space by analysing causal paths

In our case, a *causal path* links those cells that may share information derived from a common ancestor, i.e. it goes backward in time to potential sources of information and then traces possible alternate routes to present cell objects (as indicated by blue and red paths in Fig.2).

Formally, we define the matrix of causal paths between cells at a certain time-point τ as

$$B^{(\tau)} = \sum_{k=1}^{\tau} \Big(\prod_{t=\tau}^{\tau-k} (A^{(t)})^T \cdot \prod_{t=\tau-k}^{\tau} A^{(t)} \Big), \qquad (4)$$ [1]

where the latter product describes *backward-time* paths to all cells who are potential sources of information within k time-steps, while the former product computes the respective *forward-time* paths to all cells that could have potentially be influenced by the same sources within k steps. Thus a positive value B^{τ}_{ij} indicates that cells i and j potentially share information derived from a common ancestor. Since cell division events have an edge weight of 0.5, a causal path between two cells is weighted less after more divisions.

We use the matrix of causal paths $B^{(\tau)}$ to compute the mutual information for the empirical co-occurrence of cellular states. For the example of two cell states X and Y (illustrated with black and white in Fig.2) the empirical co-occurrence is defined by the number of weighted causal paths connecting cells in these respective states. Samples of paths of increasing length between cell states are depicted for a cartoon example in Fig.2. We define a *state space matrix* D by summing up all causal paths within the groups X and Y (of black and white cells, respectively) and between these groups:

$$D = \begin{pmatrix} \sum_{i,j \in X} B^{(\tau)}_{ij} & \sum_{i \in X, j \in Y} B^{(\tau)}_{ij} \\ \sum_{i \in X, j \in Y} B^{(\tau)}_{ij} & \sum_{i,j \in Y} B^{(\tau)}_{ij} \end{pmatrix}. \qquad (5)$$

In order to quantify the clustering of cell states within the genealogies we use the mutual information computed from the normalised matrix D^*:

$$MI(D^*) = \sum_{ij} D^*_{ij} \; log \left(\frac{D^*_{ij}}{\sum_i D^*_{ij} \sum_j D^*_{ij}} \right)$$

This mutual information equals 0 if the paths are distributed independent of cell state and reaches its maximum if there are only paths between cells of the same state.

3.3. Permutation tests

The absolute value of the mutual information itself is not informative since it depends on other topological characteristics such as the depth of the genealogies, cf.[10]. Therefore, we propose to use permutation tests to check whether the obtained mutual information is significantly increased compared to what is expected at random. The respective null distribution can be estimated by shuffling the rows of the temporal adjacency matrix. Thus, we can randomise the distribution of cell states while preserving topological characteristics of the network.

[1] The index t of the former product decrements from τ to $\tau - k$.

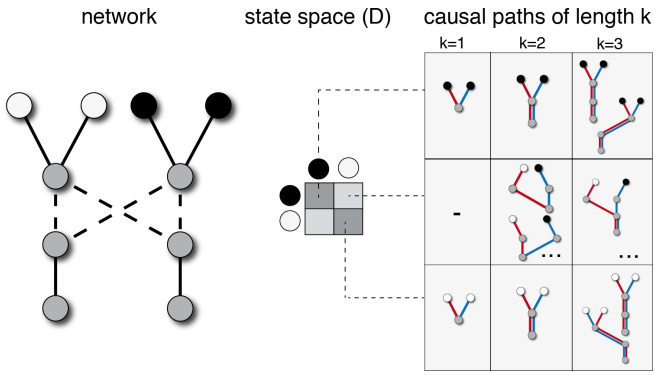


Fig. 2. **Causal paths in state space:** Selected causal paths with increasing length connecting different cell states for the example network.

4. EXPERIMENTS

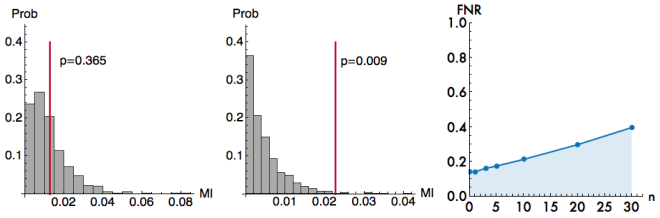
The dichotomy of *selective* vs. *instructive* regulation is a fundamental question in stem cell biology. Under certain experimental conditions, bipotent progenitor cells can give rise to progeny of a preferred lineage. There are two conceptually different paradigms of how lineage specification is regulated: either by cell survival signals for cell types committed to a certain lineage (selective), or by an intrinsic bias in lineage choice during development (instructive).

To demonstrate that our network approach is suitable to disentangle the underlying regulatory mechanism we use simulations obtained by a simple agent-based model where cellular movement is modelled as a random walk and cell differentiation into two possible lineages is described by a Pólya urn model as described in [12]. We modelled two scenarios of regulatory mechanisms underlying cellular differentiation: either regulated by an *instructive* mechanism where cells are more likely to differentiate into one particular lineage, or a *selective* mechanism, where cells differentiate with equal probability into either of the two lineages, but cells of a certain type are more likely to undergo apoptosis. Phenomenologically, both simulated scenarios result in an over-representation of one lineage. We simulated artificial movies, that allow to define possible sources of tracking errors. For each scenario, we simulated 100 movies, each with 10 initial undifferentiated cells leading to around 50 to 200 cells after 1000 time steps.

5. RESULTS AND DISCUSSION

Using the methods introduced above we can distinguish between both scenarios based on the topology of the temporal networks, *without having further information on the lineage differentiation* of the cells. Fig.3 shows representative results of the permutation test procedure: Since cell death events occur randomly in an instructive scenario, the mutual informa-

tion is consistent with the null distribution (Fig.3a). In a selective regulation scenario however, cell death is more likely to occur within one lineage and is thus clustered within certain genealogical branches leading to a significantly increased mutual information compared to the expected null distribution (Fig.3b). The fraction of instructive scenarios that erroneously led to a significant results was below 5% (false positive rate) whereas the fraction of selective scenarios with a nonsignificant result was 15% (false negative rate).



(a) instructive (b) selective (c) false negatives

Fig. 3. **Results:** (a, b) Permutation tests discriminate between regulatory regimes in simulation experiments. Estimated null distribution shown as histograms, actual result as red line. (c) False negative rate of method with respect to number of ambiguous assignments.

To test the robustness of the presented framework with respect to tracking errors, we artificially inserted increasing numbers of ambiguous assignments (50:50 chance of switching cell tracks) into our simulated data (see Fig.2). The false negative rate increases linearly with the number of ambiguities. But, even for 30 tracking errors, the method was able to correctly identify more than 60% of the simulated selective scenarios (Fig.3c).

These results show that the proposed framework can be used to analyse correlations of cell states by means of potential causal relationships, even for non-perfect tracking results. We are able to reproduce established measures designed for analysing cellular genealogies and provide a statistical testing procedure to decide between possible alternatives. The spatiotemporal networks transcend the traditional notion of genealogies as they facilitate the integration of uncertainty. Furthermore, this framework can explicitly account for potential spatial cell-cell communication (mediated by contacts or short-range soluble factors). Since imaging methods are the prime tool for assessing organisation of multicellular systems in space and time, this network approach is a valuable extension of traditional data structures.

6. REFERENCES

[1] M.A. Rieger, P.S. Hoppe, B.M. Smejkal, A.C. Eitelhuber, and T. Schroeder, "Hematopoietic Cytokines Can Instruct Lineage Choice," *Science*, vol. 325, no. 5937, pp. 217–218, July 2009.

[2] T. Schroeder, "Long-term single-cell imaging of mammalian stem cells," *Nat Methods*, vol. 8, no. 4s, pp. S30–S35, Mar. 2011.

[3] R. Tomer, K. Khairy, F. Amat, and P.J. Keller, "Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy," *Nat Methods*, vol. 9, no. 7, pp. 755–763, June 2012.

[4] E. Meijering, O. Dzyubachyk, and I. Smal, "Methods for Cell and Particle Tracking," *Imaging and Spectroscopic Analysis of Living Cells*, p. 183, 2012.

[5] R. Schnabel, H. Hutter, D. Moerman, and H. Schnabel, "Assessing normal embryogenesis in Caenorhabditis elegans using a 4D microscope: variability of development and regional specification.," *Dev Biol*, vol. 184, no. 2, pp. 234–265, Apr. 1997.

[6] O. Al-Kofahi, R.J. Radke, S.K. Goderie, Q. Shen, S. Temple, and B. Roysam, "Automated cell lineage construction: a rapid method to analyze clonal development established with murine neural progenitor cells.," *Cell cycle*, vol. 5, no. 3, pp. 327–335, Feb. 2006.

[7] D.H. Rapoport, T. Becker, A. Madany Mamlouk, S. Schicktanz, and C. Kruse, "A Novel Validation Algorithm Allows for Automated Cell Tracking and the Extraction of Biologically Meaningful Parameters," *PLoS ONE*, vol. 6, no. 11, pp. e27315, Nov. 2011.

[8] R.E. Nordon, K. Ko, R. Odell, and T. Schroeder, "Multitype branching models to describe cell differentiation programs.," *J Theor Biol*, vol. 277, no. 1, pp. 7–18, May 2011.

[9] B. Dykstra, J. Ramunas, D. Kent, L. McCaffrey, E. Szumsky, L. Kelly, K. Farn, A. Blaylock, C. Eaves, and E. Jervis, "High-resolution video monitoring of hematopoietic stem cells cultured in single-cell arrays identifies new features of self-renewal," *P NATL ACAD SCI USA*, vol. 103, no. 21, pp. 8185–8190, 2006.

[10] I. Glauche, R. Lorenz, D. Hasenclever, and I. Roeder, "A novel view on stem cell development: analysing the shape of cellular genealogies," *Cell Proliferat*, vol. 42, no. 2, pp. 248–263, Apr. 2009.

[11] C. Marr, M. Strasser, M. Schwarzfischer, T. Schroeder, and F.J. Theis, "Multi-scale modeling of GMP differentiation based on single-cell genealogies," *FEBS Journal*, vol. 279, no. 18, pp. 3488–3500, July 2012.

[12] I. Glauche, M. Cross, M. Loeffler, and I. Roeder, "Lineage specification of hematopoietic stem cells: mathematical modeling and biological implications," *Stem Cells*, vol. 25, no. 7, pp. 1791–1799, 2007.