Computational prediction of neural progenitor cell fates

Andrew R. Cohen[1, 2, 7], Francisco L.A. F. Gomes[3, 4, 7], Badrinath Roysam[2, †], Michel Cayouette[3, 5, 6, †]

[1] Department of Electrical Engineering and Computer Science, University of Wisconsin - Milwaukee, WI, USA

[2] Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

[3] Cellular Neurobiology Research Unit, Institut de recherches cliniques de Montréal, Montréal, QC, Canada

[4] Department of Biology, Universidade Estadual do Ceará, Fortaleza, Ce, Brazil

[5] Department of Medicine, Université de Montréal, Montréal, QC, Canada

[6] Departments of Biology and Anatomy and Cell Biology, McGill University, Montreal, QC, Canada

[7] Authors contributed equally to this work

[†] Authors for correspondence (roysam@ecse.rpi.edu; michel.cayouette@ircm.qc.ca)

Author contributions:

A.C. and B.R. developed the computational methods described here. F.G. and M.C. proposed the initial hypothesis and developed the cell culture, immunostaining, and imaging methods described here. All authors contributed to writing the manuscript.

Key Words: retina, self-renewal, stem cell, neural development, cell-fate decision, cell-fate choice, computational biology, algorithmic information theory

ABSTRACT

Understanding how stem and progenitor cells choose between alternative cell fates is a major challenge in developmental biology. Efforts to tackle this problem have been hampered by the scarcity of markers that can predict cell division outcomes. Here we present a computational method based on algorithmic information theory that can analyze dynamic features of living cells over time. Using this method, we asked whether rat retinal progenitor cells (RPCs) display characteristic phenotypes before undergoing mitosis that could foretell their fate. We were able to predict whether RPCs will undergo a self-renewing or terminal division with 99% accuracy, or whether they will produce two photoreceptors or another combination of offspring with 87% accuracy. Our implementation can segment, track and generate predictions for 40 cells simultaneously on a standard PC at 5 minutes/frame. This method could be used to isolate cell populations with specific developmental potential, thus permitting previously impossible investigations.

INTRODUCTION

In the developing vertebrate nervous system, hundreds of different types of neurons and glial cells are produced from multipotent neural stem/progenitor cells. How is such tremendous diversity generated? The retina is a convenient model to study cell fate determination mechanisms in the central nervous system. It has a laminated structure containing only seven different cell types that are all identifiable by specific markers. These cell types are generated from a pool of multipotent retinal progenitor cells (RPCs) in a strict, but overlapping chronological order of cell birth during retinal development. Although RPC fate decisions are influenced by feedback inhibition signals, growing evidence suggests that cell autonomous developmental programs also have a critical regulatory role[1]. We and others proposed that lineage-dependent cell fate decisions may be "pre-programmed" into individual neural progenitor cells to generate the correct combination of cell types at specific times[2-5]. If this model were correct, one would expect a large heterogeneity in the progenitor cell population. Gene profiling experiments performed on whole RPC populations at different stages of development identified potential regulators of cell fate[6-9], but did not provide information about heterogeneity of the RPC population. Recently, however, single-cell gene expression profiling revealed a large heterogeneity in gene expression among mouse RPCs[10], even at the same stage of development, suggesting that different populations of RPCs with specific developmental potential might exist at any one stage of development. Heterogeneity in gene expression among a progenitor population is most likely not unique to the retina. For example, single-cell gene expression profiling of olfactory progenitor cells also revealed transcriptional differences among the progenitor cell population[11], and it is well known that spinal cord development depends on the diversification of the progenitor cell pool[12]. Together, these results suggest that a given population of seemingly indistinguishable progenitor cells might actually be composed of many different sub-populations of fate-restricted progenitor cells. Studying this heterogeneity will require a method that can predict the fate outcome of an individual progenitor *before* it divides. This would provide invaluable information when interpreting results of single cell gene profiling experiments, and could help identify novel markers of sub-populations of progenitor cells with specific developmental potential.

We hypothesized that distinct gene expression profiles in RPCs with specific developmental potential might translate into characteristic dynamic behaviors. These behaviors are likely to be subtly different, and difficult or impossible for a human observer to discern visually. Therefore, testing this hypothesis requires automated image analysis tools capable of accurately identifying differences in populations based on dynamic behaviors. In this paper, we present a method named Algorithmic Information Theoretic Prediction (AITP) that combines recent advances in image sequence summarization[13] and semi- supervised spectral learning[14] to develop this capability.

In this paper, we present a method named Algorithmic Information Theoretic Prediction (AITP) that combines recent advances in image sequence summarization[13] and semi- supervised spectral learning[14] to develop this capability. AITP provides a sensitive parameter-free multiresolution analysis of the multi-dimensional time-sequence data representing the observable spatio-temporal dynamic phenotype of cells obtained from segmentation and tracking algorithms. AITP achieves a multiresolution analysis of the dynamic phenotype by analyzing the multidimensional numeric time series data at multiple levels of quantization and automatically selecting the optimal resolution.

Our results indicate that RPCs display distinctive behavior, even before they undergo mitosis, depending on which combination of daughter cell types they are going to generate. In addition, we report that AITP can be used to predict cell fate outcomes of oligodendrocyte precursor cells (OPCs) as well. Combined with single cell microarray technology, this method should prove invaluable in the future to identify molecular markers characteristic of specific populations of various stem/progenitor cells. All of the AITP software can be downloaded at: https://pantherfile.uwm.edu/cohena/www/aitpd%20main.html

RESULTS

Long-term time-lapse microscopy of RPCs

To determine whether RPCs display characteristic dynamic patterns that can predict their fate, we first developed a way to image live RPCs as they divide and give rise to various retinal cell types over a period of several days. To do this, we cultured embryonic day 20 rat RPCs at clonal density, as previously described[2]. Four hours after plating, the culture dish was placed in a temperature and CO_2-controlled environment chamber fitted around an inverted microscope. Presumed RPCs were then identified by their characteristic neuroepithelial morphology. Using a motorized microscope stage, >100 cells were selected in each experiment and imaged under phase-contrast microscopy every 5 minutes over a period of 9-13 days.

At the end of the recording, we used a combination of cell morphology and expression of cell-type specific markers to reliably identify the four different retinal cell types produced from E20 RPCs, as we previously reported[2] (**Fig. 1**). The outcome of the first division of each RPC was retrospectively determined by playing back the time-lapse movie and tracking the fate of both daughter cells. Some daughter cells kept an RPC morphology and divided again, indicating that they remained RPCs. Some others did not divide again and changed morphology, indicating that they differentiated. In this latter case, retinal cell types were identified as summarized in **Supplementary Table 1**. Overall, we could reliably identify the outcome of RPC divisions as either 1) proliferative, generating two RPCs (not shown); 2) self-renewing, generating a RPC and a neuron or glial cell (**Fig. 2; Supplementary Movie 1**); or 3) terminally differentiating, generating two neurons of the same or different types (**Fig. 3; Supplementary Movie 2**).

Automated measurement of cell morphology and dynamic phenotype

The RPCs of interest were segmented (computationally delineated) and tracked over time (**Fig. 4**). An example of segmentation and tracking results for one RPC image sequence is provided online (**Supplementary Movie 3**). The segmentation and tracking results provide the basis for computing time courses of cell features that quantify the dynamic cellular phenotypes. The segmentation and tracking are application dependent tasks that provide the input for the *application independent* AITP methodology; the AITP method can be applied directly to any application for which segmentation and tracking algorithms are available. The location of a cell, denoted $(x(t), y(t))$ is estimated as the median x and y coordinates of all pixels constituting the cell in the image. Using these locations, we compute cell movement vectors $(\Delta x(t), \Delta y(t))$, , and movement directions $\theta(t)=tan^{-1}(\Delta y(t)/\Delta x(t))$. We also compute the net movement of each cell from its initial location, denoted $D(t)= |(x(t),y(t))- (x(0),y(0))|$, as previously reported[15]. Next, we fit an ellipse to each cell region, and compute its eccentricity $e(t)$. Finally, the size of the cell $S(t)$ is computed as the area of its convex hull. These calculations yield a time sequence of 6-dimensional feature vectors $\mathbf{F}=[\Delta x(t), \Delta y(t), \theta(t), D(t), e(t), S(t)]$, one for each cell, for subsequent analysis. This set of features represents the simplest choices. Additional features could be included in the future. The included features are not necessarily independent of each other; using derived features (*e.g.* $\theta(t)$, $D(t)$) can improve the accuracy of the compression based distance measure used in the subsequent analysis. At this stage, it is unimportant to be concerned about choosing the *optimal* set of features since irrelevant features are identified and discarded automatically by a later computational step.

Algorithmic information theoretic analysis of the feature vectors

The AITP method is a general-purpose approach to analyzing the multidimensional time sequence data obtained from image segmentation and multitarget tracking algorithms. The dynamic cellular phenotypes, as captured by the time sequences of feature vectors, were compared using the normalized compression distance measure (NCD). This measure was reported recently in the field of algorithmic information theory[16], and was enhanced by us to include automatic quantization (see details in the **Methods** section). Briefly, the NCD is used to calculate a $M \times M$ pair-wise distances matrix $\mathbf{A}$, where M is the number of RPCs. The $(i,j)^{\text{th}}$ element of this matrix is the NCD between the features of RPCs i and j. This calculation is repeated for multiple quantization levels N=1..$N_{\max}$, and for each element f of the feature vector power set. The special value N=1 corresponds to no quantization. The above calculations resulted in $N_{\max} \times (2^{|F|}-1)$ distance matrices, where $|F|$ is the number of features. If this number exceeds available computational capacity, feature subset selection methods[13]can be used to reduce the computational burden.

This collection of distance matrices is analyzed using a cross-validated semi-supervised spectral learning methodology[14] to assign predictions to cells with unknown outcomes by comparing their dynamic behaviors to cells from the training data that have known outcomes (mathematical details are presented in the **Methods** section, and the software can be downloaded at: https://pantherfile.uwm.edu/cohena/www/aitpd%20main.html). A numeric value is assigned to each possible outcome that we want to predict based on dynamic behavior. For a 2-class problem , *e.g.* whether a RPC will undergo a terminal versus a self-renewing division or whether a terminal RPC will produce two photoreceptor neurons, versus another combination of offspring (**Table 1**, top and middle rows), the outcome for an RPC is a 0 or 1. For a 3-class problem *e.g.*

whether a RPC will produce two photoreceptors versus a photoreceptor - bipolar pair or a photoreceptor - amacrine pair (**Table 1**, bottom row), it is 0, 1, or 2.

Prediction of self-renewing divisions

To determine whether AITP could reliably predict self-renewing vs. terminal divisions, we analyzed the time sequence data of all RPCs for which we could reliably determine the outcome of the first division. The results are shown in a confusion matrix (**Table 1**). Remarkably, only one out of nineteen self-renewing RPCs was erroneously predicted to be a terminal RPC, and the outcome for all fifty-three terminal RPCs was correctly predicted, corresponding to an accuracy of 99%. The 95% confidence interval for this prediction is 92.5%-99.8%[17]. A minimum of one hundred image frames (about 8 hours of recording) was chosen arbitrarily for cells to be included in the AITP analysis. These results indicate that RPCs exhibit distinctive dynamic behaviors depending on whether they will undergo a terminal or a self-renewing division and that this pattern can be recognized using AITP.

In the above experiments, the entire dataset was used for semi-supervised training by partitioning the data using the method of leave-one-out cross validation. This raises the question of whether training data from one experiment are applicable to an independent experiment. To address this question, we took advantage of the fact that the data for the 72 cells analyzed in **Table 1** originated from three independent experiments (see **Supplementary Table 2**). When we analyzed the cells from Experiment 3 using only the outcomes from Experiments 1 and 2 for training, we found that only one cell out of the 23 was misclassified, giving a prediction accuracy of 96% (**Supplementary Table 3**). These results indicate that training data gained from one experiment are applicable to another independent experiment, and suggest that slight variations in culture conditions from one experiment to another should not affect the quality of the results.

We next asked whether AITP could be applied to predict cell fate outcomes of a different neural progenitor cell type. To address this question, we cultured oligodendrocyte precursor cells (OPCs), which generate oligodendrocytes, the myelinating cells of the central nervous system. These cells, when cultured under differentiating conditions (see **Methods**), divide to generate two oligodendrocytes (terminal division; **Supplementary Fig. 1 and Movie 4**), one OPC and one oligodendrocyte (self-renewing division; **Supplementary Fig. 2 and Movie 5**), or two OPCs (proliferative division; not shown). The cells were imaged for 7-10 days, as described for RPCs, and AITP was applied on the resulting movies. We were able to predict whether OPCs would undergo a terminal or a non-terminal division with 88% accuracy. These results suggest that AITP could potentially be applied to different progenitor cell types, provided that the cells can be tracked and segmented correctly.

Prediction of cell fate outcome of terminal divisions

Since the majority of terminal divisions in the neonatal rat retina produce two photoreceptor cells, we next asked whether we could use the AITP method to discriminate between terminal divisions generating two photoreceptors (Ph/Ph) or any other combinations of offspring. By analyzing RPCs that will undergo a terminal division, we found that AITP could predict with 87% accuracy whether the outcome of the division would be a daughter cell pair containing two photoreceptors or some other combination of cell types (**Table 1b**). The 95% confidence interval for this prediction is 78.5%-92.7%[17]. More precisely, RPC terminal divisions generated two photoreceptors, a photoreceptor/bipolar neuron, or photoreceptors/amacrine neuron cell pair, and AITP could predict the actual outcome of such divisions with 83% accuracy (**Table 1c**). These results indicate that AITP can not only predict self-renewing vs. terminal divisions, but also whether a RPC will produce a particular combination of neurons.

Live prediction of cell fate

In the experiments above, the fate predictions were achieved retrospectively, after the cells were fixed and identified by immunocytochemistry. This was necessary in the first stages of development of AITP as it provided a way to cross validate the predictions. However, a “live” fate prediction functionality for AITP would allow one to isolate cells predicted to have a particular fate during the course of the experiment, before they divide. To add this functionality to AITP, we modified and optimized the algorithm to run at a much faster rate, as described in the **Methods** and provided at: https://pantherfile.uwm.edu/cohena/www/aitpd%20main.html. To test the real-time capability of this new algorithm, we used the time-lapse microscopy data that we had generated, and simulated live acquisition by making one image available every 5 minutes, starting from the first frame of the movie until the last frame before mitosis. We were able to segment, track and generate predictions for 40 cells simultaneously on a standard PC (dual quad core Intel Xeon X5472 3GHz processors, 8GB RAM, Windows Vista) within the five-minute per frame microscope acquisition time. This data required approximately 45 seconds per frame to segment and update the tracking, and an additional three to five seconds per cell to generate a prediction. The computation is inherently parallel; more cells can be analyzed simultaneously by adding more processors. Using this algorithm, we obtained the same fate prediction accuracy as in the retrospective analysis, confirming the reliability of this approach.

DISCUSSION

The value of our method lies in its ability to make up for the difficulty of visually identifying subtle shape and movement pattern differences. Taken together, our results suggest that RPCs in culture display dynamic patterns that can be sensed computationally to predict the outcome of their next division using the new generation of analytic tools represented by algorithmic statistics and algorithmic information theory. We discuss some of the implications of this method below.

One exciting potential application opened up by this method is the possibility to purify a population of RPCs, or any other cell type that can be imaged, with a specific developmental potential. Although single-cell gene chip experiments have shown a high degree of heterogeneity of gene expression among the RPC population[10], these results are difficult to interpret because the fate of any particular RPC was unknown. By applying the method presented here on live RPCs (or another type of progenitor cell), it would be possible, for example, to isolate RPCs that would have undergone a self-renewing division at their next mitosis and compare their gene expression profiles using single-cell microarray to that of RPCs that would have undergone a terminal division. Such experiments could help identify genes involved in self-renewing divisions, a mode of cell division highly relevant to stem cell biology. As the method can also predict the outcome of terminal divisions with high accuracy, it should be possible to isolate RPCs with different differentiation potential, which might lead to the identification of novel genes involved in the specification of various retinal cell types. The cell cycle of neonatal RPCs is about 36-40 hours, and we have shown that 8 hours of recording is sufficient to generate an accurate prediction. Since the majority of RPCs would not undergo mitosis in the first 8 hours of recording, it should be possible to generate a computational prediction on living RPCs and isolate them before they undergo mitosis.

To adapt this method to applications with other cell types or using a broader set of cell features requires merely the availability of effective cell segmentation and/or tracking algorithms, and a set of relevant features. Such algorithms are increasingly available for 2-D, 3-D and multi-channel time-lapse data[18]. As proof-of-principle that AITP can be extended to other cell types, we show that AITP can be used to accurately predict cell division pattern of OPCs, a different type of neural progenitor cells. Figure 4 illustrates the automated cell segmentation and tracking approach used for the image sequence data analyzed here. Although both cell types analyzed in this study were cultured at low density, it should be possible to apply AITP to cells cultured under high-density conditions as segmentation and tracking algorithms robust enough to process hundreds of such image sequences with a minimum of user interaction become available.

All subsequent procedures for analyzing the feature data are independent of the particular cell type or application, and represent a common set of tools enabled by the recent advances in algorithmic information theory and algorithmic statistics. The method automatically identifies the feature subset that gives the best prediction of progenitor fate. Identifying the specific behaviors within the feature subset that differentiate the populations of interest is a topic for future research. From an algorithmic information theoretic standpoint, we are estimating a normalized form of the length in bits of the Universal Turing machine program to convert between time sequence data; this is a different problem from actually identifying the specific program. From a practical standpoint, efforts to understand what similarities the real world compression algorithm is identifying in the multidimensional time sequence data have not provided any useable insight.

One practical issue to consider is the sheer volume of time-lapse image data, and the computational complexity of the analysis. Accordingly, our software is written to take advantage of any parallel computing facility that supports the common MPI library. We also provide a version that runs on serial computers for generality. We have also implemented a highly optimized version of our prediction algorithm that enables the software to run live as the cells are being imaged on the microscope, providing the capability to collect the cells of interest without stopping the time-lapse acquisition. When we tested this algorithm, it obtained the same fate prediction accuracy as in the retrospective analysis, confirming its reliability.

In conclusion, our findings suggest that specific cellular phenotypes can be recognized by carefully analyzing dynamic cellular features. As AITP could be applied to the analysis of any cell type of interest that can be imaged, segmented and tracked over time, it could potentially have applications in various fields.

METHODS

Retinal progenitor cell culture, immunostaining, and time-lapse imaging

All animal experiments were done in accordance with the Canadian Council on Animal Care and approved by the Institut de Recherches Cliniques de Montreal animal care committee. Retinal cells from embryonic day 20 (E20) Sprague Dawley rat retinas were cultured as previously described[2]. Approximately 2 x 10^4 cells were plated in 35 mm Falcon dishes coated with poly-L-lysine (10 μm/ml) and laminin (10 μg/ml). The dissociated cells were allowed to settle for four hours in a CO_2 incubator at 37° C before they were placed under the time-lapse microscope. RPCs were imaged with a Leica 20X phase contrast objective and the images captured using a Hamamatsu CCD video camera connected to a Macintosh computer equipped with Volocity software (Improvision) programmed to capture a frame every 5 minutes. The cells were kept at 37ºC, 8% CO_2 and 12% O_2. The reduced level of O_2 was achieved by pumping N_2 into the chamber.

The retinal cells were fixed after 9-13 days and the following antibodies were used for immunofluorescence: monoclonal mouse anti-islet-1 (1:2000; produced by T. Jessell and obtained from the Developmental Studies Hybridoma Bank) and rabbit anti-Pax-6 (1:10000; Santa Cruz Biotech.). Primary antibodies were detected using goat anti mouse IGg2b Alexa Fluor-488 and goat anti rabbit IgG Alexa Fluor-594 (Molecular Probes). In all cases, we counterstained the nuclei with Hoechst 33342 (Molecular Probes). Three independent experiments were performed to generate the entire dataset.

Oligodendrocyte precursor cell (OPC) culture and time-lapse imaging

For each experiment, at least twenty optic nerves were removed from postnatal day 7 (P7) Sprague Dawley rats and dissociated and cultured as described previously[19]. Prior to time-lapse imaging half the medium was replaced with MBS medium supplemented with T3 (Thyroid hormone, Sigma, 40 ng/ml) to induce OPC differentiation [20]. Time-lapse setting was the same as for RPC imaging (see above). Cells were identified based on their characteristic morphology. Three independent experiments were performed to generate the entire dataset.

Cell Segmentation and Tracking in Time-Lapse Phase Contrast Data

Cells in phase-contrast exhibit a characteristic halo that we exploit for segmentation. Pixels whose brightness is five standard deviations above the average pixel intensity are considered halos or bright interior pixels. Morphological opening and closing operations[21] are used to smooth the halo regions, and to identify a bounding region for each cell by filling any holes inside the halos (**Fig. 4b**). Halo pixels are then separated into connected regions using the watershed transform[22]. The morphological smoothing prevents unwanted fragmentation[23]. A spectral *k*-means clustering[24] of the pixel intensities within the bounding region for each cell groups the pixels into three groups. The brightest and darkest groups belong to the cell (green and yellow pixels in **Fig. 4c**), and the intermediate-intensity halo pixels (red) are discarded. A constrained watershed transform is applied to the cell pixels to separate touching cells. The watershed transform is constrained using the H-maxima transform[21] to reduce fragmentation.

To track cells, we use a variation of the method described previously [25]. Importantly, we use tracking anomalies to identify scenarios where the watershed transform has resulted in erroneous cell fragmentation , and then merge the over-segmented regions automatically. This is implemented using a temporal look-ahead step whenever the assignment matrix used by the tracking algorithm is not square. In that case, our algorithm checks for temporal consistency up to

9 frames ahead in the video sequence to see if merging adjacent cell regions results in more consistent tracking (**Fig. 4g**). If so, the regions are merged (**Fig. 4h**). This results in a time series of cell outlines (**Fig. 4i**). During the tracking, dividing cells are detected automatically. When a cell divides, the nuclei of daughter cells appear distinctly in phase contrast as bright regions of approximately equal size. Using morphological opening and connected component analysis, we examine whether there are two such bright regions. This is illustrated in panels **j** and **k** of **Fig. 4**. The first time that two such regions are detected is considered the time of mitosis (panel **k**). Prior to applying AITP, the results are screened visually for tracking errors. Cells with tracking errors are excluded from further analysis.

Importantly, there are a wide range of approaches to segmenting and tracking different cell types imaged using various techniques[26-29]. AITP can be applied to any application for which segmentation and tracking algorithms are available.

Computation of the quantization-enhanced NCD distance Matrix

The NCD distance measure was proposed as a practical tool for analyzing unstructured data[16]. The theoretically ideal distance measure for comparing arbitrary datasets is the normalized information distance[30] based on Kolmogorov's algorithmic information theory[31] that can account for *any and all* differences. However, this measure is impractical due to its uncomputability[32]. To overcome this limitation, the normalized compression distance (NCD) was proposed as a practical approximation to the normalized information distance using data compression algorithms[16]. To use this measure, the feature vectors are treated as symbol strings. Let $C(x)$ denote the size in bytes of the compressed version of string x, and $C(x;y)$ the size of the compressed version of the concatenation of x and y. Then,

$$NCD(x,y) = \frac{C(x;y) - \min(C(x),C(y))}{\max(C(x),C(y))}. \qquad (1)$$

The NCD is unique in its ability to accurately compare multidimensional data sequences of varying length in a parameter-free manner. Problem formulations based on NCD can be general, parameter free, robust to noise, and independent of applications and data formats [16,33,34]. They can also overcome the main practical limitation of Minimum Description Length (MDL)[35,36] based techniques for approximating Kolmogorov complexity – the need for "tightly-tuned" application-specific models and data representations. There are no parameters needed to compute the NCD, except for the choice of compression algorithm. It was shown that the choice of compression algorithm has a negligible impact on the final analysis[16]. We used the `bzip2` compressor. Previously, we showed that the accuracy of the NCD can be enhanced further by *quantizing* the time sequence data using N symbols, where the optimal value of N can be estimated automatically [13]. This procedure enables a multi-resolution analysis since more symbols imply a finer resolution and *vice versa*. In our examples, we used a maximum ($N_{\max}$) of 26 symbols, corresponding to the letters of the alphabet (typically used by text-based compressors). We limited ourselves to atomic (single) symbols to ensure that the symbols themselves do not introduce any artifactual similarities in the data. The goal of the quantization is to assign symbols to numeric data such that all symbols are equiprobable. Source code for automatically quantizing the time-sequence data and computing the NCD are included at: https://pantherfile.uwm.edu/cohena/www/aitpd%20main.html.

Method to transform the NCD distance matrices to eigenspace

Let **D** denote the diagonal matrix whose diagonal $(i,i)^{\text{th}}$ element is the sum of the i^{th} row of the NCD distance matrix **A**. This matrix is first normalized using the formula

$$\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \quad (2)$$

The normalized matrix **L** is symmetrical with zeros on the diagonal (i.e., it is positive semi-definite). The eigenvectors of **L** are sorted by the magnitudes of the eigenvalues, and used to form the columns of a new $M \times M$ matrix denoted **V**. This is known as the spectral matrix because it is based on an eigenspace representation. The i^{th} row of this matrix is an eigenspace representation of the i^{th} RPC. The k left-most columns of the spectral matrix correspond to the k principal eigenvectors. When just these k columns of the i^{th} row are retained, we obtain a reduced k-dimensional representation of the entire time-course of features for the i^{th} RPC. We use the symbol $v_i^{k,N,f}$ to denote this k-dimensional point. The above data transformation into eigenspace is profoundly valuable because it casts the multi-dimensional time-sequence data in a form that naturally enables the use of unsupervised and/or supervised machine learning algorithms (clustering, nearest neighbors, decision trees, *etc*.).

Semi-supervised analysis of the L-matrix with cross validation

Unsupervised analysis methods, as previously described[13], are appropriate for exploring unlabeled data, i.e., when the outcomes of cell division are not provided to the computational method. When the fates of some of the RPCs are known, these data can be provided as labeled training examples to a supervised learning method. Classic supervised methods learn from these examples, and then process the unlabeled data. Recently, a new class of *semi-supervised* data analysis methods have emerged that can exploit the unlabeled data in addition to the training data for improved learning. In this approach, both labeled and unlabeled data are used in the transformation to eigenspace, so the unlabeled data influence the spectral matrix via the eigenvalues and eigenvectors of the **L** matrix. A supervised learning algorithm that still uses only the labeled data points for training benefits from the improved spectral eigenvalues. Such algorithms have been shown to outperform classical supervised learning[14] methods. This is the method adopted by us.

We used the following semi-supervised learning with a leave-one-out cross validation. For each value of k from 1 to M, we applied supervised classification algorithms, denoted C^j to predict RPC fates. The prediction for the i^{th} RPC is denoted $C^j(v_i^{k,N,f})$. For a 2-class problem (*e.g.*, **Table** 1, top and middle rows), this is a 0 or 1. For a 3-class problem (*e.g.*, **Table 1**, bottom row), it is 0, 1, or 2.

Now, for the RPC i with known fate y_i, we can compare our prediction $C^j(v_i^{k,N,f})$ based solely on the features (not including RPC i in the training data) to y_i. This is the method of leave-one-out cross validation[17]. The validation error for classifier C^j is estimated as:

$$\varepsilon_{C^j}^{k,N,f} = \frac{\sum_{i=1}^{M} H^j(v_i^{k,N,f})}{M}, \text{ where } H^j(v_i^{k,N,f}) = \begin{cases} 0 & C^j(v_i^{k,N,f}) = y_i \\ 1 & \mathit{otherwise} \end{cases}. \tag{3}$$

Using the above procedure, we identify the values that minimize the validation error,

$$\hat{f},\hat{k},\hat{N},\hat{C} = \min_{f,k,N,C^j} \varepsilon_{C^j}^{k,N,f} \tag{4}$$

. Software source code written in C (both single processor and MPI versions) for performing the above analysis is provided at: https://pantherfile.uwm.edu/cohena/www/aitpd%20main.html

Simulated live prediction of cell fate outcome

Live cell fate prediction begins with the segmentation and tracking results for a population of cells with known post-mitotic outcomes. Prior to applying live cell fate prediction, AITP is used to determine the parameters that give the most accurate prediction of cell fate for this population of cells. There are two components in the live cell fate prediction algorithm. The segmentation and tracking component starts a process for each live cell that we wish to analyze. This process attaches to an individual folder on a computer connected to the microscope. When a new image frame appears, the process wakes up, segments the new frame, updates the tracking and then waits for the next frame. The second component takes the current segmentation and tracking results generated for an individual cell and generates a prediction of that cell's fate using the settings obtained from the population of cells with known outcomes.

ACKNOWLEDGMENTS

The computational aspects of this work were supported by the Center for Subsurface Sensing and Imaging Systems (NSF Grant EEC-9986821), by the Rensselaer Polytechnic Institute and by the University of Wisconsin-Milwaukee. We thank Adam Todorski and the staff at Rensselaer's CCNI supercomputing center. We thank Sally Temple for helpful comments and feedback, and on collaborations leading up to the present work. We thank Marie Kmita and Jonah Chan for insightful comments on the manuscript, Aude Daigneault for expert technical assistance with the animal colony, and Christine Jolicoeur for help with the oligodendrocyte precursor cell culture. This work was supported by grants from the Canadian Institutes of Health Research and the Foundation Fighting Blindness – Canada (to M.C). M.C. is a CIHR New Investigator and a W.K. Stell Scholar of the Foundation Fighting Blindness – Canada.

FIGURE LEGENDS

Figure 1. Retinal cell type identification. Cell types are identified based on morphology and immunostaining for cell type-specific markers, as indicated. (**a-d**) A clone composed of two rod photoreceptors (Ph, Pax-6 -, Islet-1-) with characteristic round cell-body, thin processes, and heterochromatin condensation. (**e-h**) A clone composed of one amacrine cell (Am, Pax-6 +, Islet-1-), one rod photoreceptor, and one bipolar cell (Bi, Pax-6 -, Islet-1+). (**i-l**) A clone composed of one amacrine cell and one bipolar cell. (**m-p**) A Muller glial cell (Mu), note the distinct glial morphology, large nucleus, absence of neurites and lack of expression of neuronal markers. Scale bars = 25µm (**a-l**), 40µm (**m-p**).

Figure 2. Self-renewing divisions. (**a**) Snapshots of time-lapse video microscopy showing a rod-committed RPC undergoing three rounds of self-renewing divisions to generate a clone containing four rod photoreceptors; dashed circles depict mitotic cells in telophase. (**b**) Lineage reconstruction of the clone showed in (**a**); arrow indicates the progenitor whose behavior was analyzed to generate prediction using AITP; (**c**) Hoechst staining of cell nuclei. Ph: photoreceptor. Time is given in h:min. Scale bar = 15µm.

Figure 3. Terminal division. (**a**) Snapshots of time-lapse video microscopy showing a RPC undergoing a terminal division to generate a clone containing one rod photoreceptor and one amacrine cell. Dashed circles depict mitotic cells in telophase. (**b**) Immunostaining for Islet-1 (green), Pax-6 (red), and Hoechst (blue). (**c**) Lineage reconstruction of the clone showed in (**a**).

Arrow points to the RPC whose behavior was analyzed using AITP; Ph, photoreceptor; Am, amacrine cell. Time is given in h:min. Scale bar = 25µm.

Figure 4. Automated cell segmentation and tracking. (**a**) Example of a RPC imaged with phase-contrast microscopy. (**b**) The cell is separated from the background based on the halo produced by phase-contrast microscopy. (**c**) Adaptive clustering separates pixels into groups - the yellow dots indicating dark interior points, green dots indicating bright interior points, and red dots indicating points on the halo. The cyan outline is the geometric convex hull of the interior points. (**d-f**) Segmentation results at 3 intermediate points in the time-lapse sequence. (**g, h**) Procedure for correcting segmentation results using the temporal consistency constraint. **(i)** Overlay of successive color-coded cell outlines superimposed on the initial image (panel **a**) to illustrate the cell tracking results. (**j, k**) Panels illustrating mitosis detection. (**l, m**) Progeny of cell division shown in phase contrast and using a fluorescent nuclear stain (Hoechst). Scale bar = 7µm (**a-h**); 5µm (**j, k**); 15µm (**l, m**); 25µm (**i**).

Table 1. Summary of performance data for prediction of RPC fate. All three outcomes were analyzed using the same software methodology. Top row shows the results of predicting a terminal versus a self-renewing division. The middle row shows the results of predicting whether a terminal RPC will produce two photoreceptor neurons, versus another combination of offspring. The bottom row shows the results of predicting whether a RPC will produce two photoreceptors versus a photoreceptor - bipolar pair or a photoreceptor - amacrine pair.

Table 1: Summary of performance data for prediction of retinal progenitor cell fate.

			Confusion matrix			Output parameters	Correct prediction rate
			Predicted Outcome				
Self renewing vs. terminal divisions (66 movies, 72 cells)			Self-renewing	Terminal		Quantization = 17 symbols Spectral matrix dimension = 4 Feature subset= (Δx, Δy, θ, S)	99%
	True Outcome	Self-renewing	18*	1			
		Terminal	0	53			
			Predicted Outcome				
Photoreceptor (Ph) vs. non-photoreceptor progeny (76 movies, 86 cells)			Non-Ph, Ph	Ph, Ph		Quantization = 12 symbols Spectral matrix dimension = 80 Feature subset= (Δx, Δy, S)	87%
	True Outcome	Non-Ph, Ph	32	2			
		Ph, Ph	9	43			
			Predicted Outcome				
Combinations of progeny: Photoreceptor (Ph), Bipolar (Bi), Amacrine (Am) (68 movies, 78 cells)			Ph,Ph	Ph,Bi	Ph,Am	Quantization = 14 symbols Spectral matrix dimension = 64 Feature subset = (D, S, e)	83%
	True Outcome	Ph,Ph	49	0	3		
		Ph,Bi	4	4	2		
		Ph,Am	4	0	12		

* Entries indicate the number of cell divisions for each category

REFERENCES

1. M. Cayouette, L. Poggi, and W. A. Harris, *Trends Neurosci* **29** (10), 563 (2006).
2. M. Cayouette, B. A. Barres, and M. Raff, *Neuron* **40** (5), 897 (2003).
3. L. Godinho, P. R. Williams, Y. Claassen et al., *Neuron* **56** (4), 597 (2007).
4. X. Mu, X. Fu, H. Sun et al., *Curr Biol* **15** (6), 525 (2005).
5. L. Poggi, M. Vitorino, I. Masai et al., *J Cell Biol* **171** (6), 991 (2005).
6. E. Diaz, Y. H. Yang, T. Ferreira et al., *Proc Natl Acad Sci U S A* **100** (9), 5491 (2003).
7. M. I. Dorrell, E. Aguilar, C. Weber et al., *Invest Ophthalmol Vis Sci* **45** (3), 1009 (2004).
8. F. J. Livesey, T. L. Young, and C. L. Cepko, *Proc Natl Acad Sci U S A* **101** (5), 1374 (2004).
9. X. Mu, S. Zhao, R. Pershad et al., *Nucleic Acids Res* **29** (24), 4983 (2001).
10. J. M. Trimarchi, M. B. Stadler, and C. L. Cepko, *PLoS ONE* **3** (2), e1588 (2008).
11. I. Tietjen, J. M. Rihel, Y. Cao et al., *Neuron* **38** (2), 161 (2003).
12. T. M. Jessell, *Nat Rev Genet* **1** (1), 20 (2000).
13. A. R. Cohen, C. S. Bjornsson, S. Temple et al., *IEEE Trans Pattern Anal Mach Intell* **31** (8), 1386 (2009).
14. S. D. Kamvar, D. Klein, and C. D. Manning, *International Joint Conference of Artificial Intelligence* (2003).
15. LM Baye and BA Link, *J Neurosci* **27** (38), 10143 (2007).
16. R. Cilibrasi and P. M. B. Vitanyi, *Information Theory, IEEE Transactions on* **51** (4), 1523 (2005).
17. Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques* Second ed. (2005).
18. Y. Chen, E. Ladi, P. Herzmark et al., *J Immunol Methods* **340** (1), 65 (2009).
19. B. A. Barres, I. K. Hart, H. S. Coles et al., *Cell* **70** (1), 31 (1992).
20. B. A. Barres, M. A. Lazar, and M. C. Raff, *Development* **120** (5), 1097 (1994).
21. Pierre Soille, *Morphological Image Analysis: Principles and Applications*, 1st ed. (Springer-Verlag, 1999).
22. Luc Vincent and Pierre Soille, *IEEE Transactions of Pattern Analysis and Machine Intelligence* **13** (6), 583 (1991).
23. J. Lin, Keogh, E., Lonardi, S. & Chiu, B, *Data Mining and Knowledge Discovery Journal.* **15** (2), 107 (2007).
24. Andrew Y. Ng, Michael Jordan, and Yair Weiss, *Advances in Neural Information Processing Systems 14* (2002).
25. Omar Al-Kofahi, Richard J. Radke, Susan K. Goderie et al., *Cell Cycle* **5** (3), 327 (2006).
26. O. Debeir, P. Van Ham, R. Kiss et al., *Medical Imaging, IEEE Transactions on* **24** (6), 697 (2005).
27. K. Jaqaman, D. Loerke, M. Mettlen et al., *Nat Methods* **5** (8), 695 (2008).
28. K. Li, E. D. Miller, M. Chen et al., *Med Image Anal* **12** (5), 546 (2008).
29. E. Meijering, I. Smal, and G. Danuser, *Signal Processing Magazine, IEEE* **23** (3), 46 (2006).
30. C. H. Bennett, P. Gacs, Li Ming et al., *Information Theory, IEEE Transactions on* **44** (4), 1407 (1998).
31. M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. (Springer Verlag, New York, 1997).

32. M. Li, X. Chen, X. Li et al., *Information Theory, IEEE Transactions on* **50** (12), 3250 (2004).

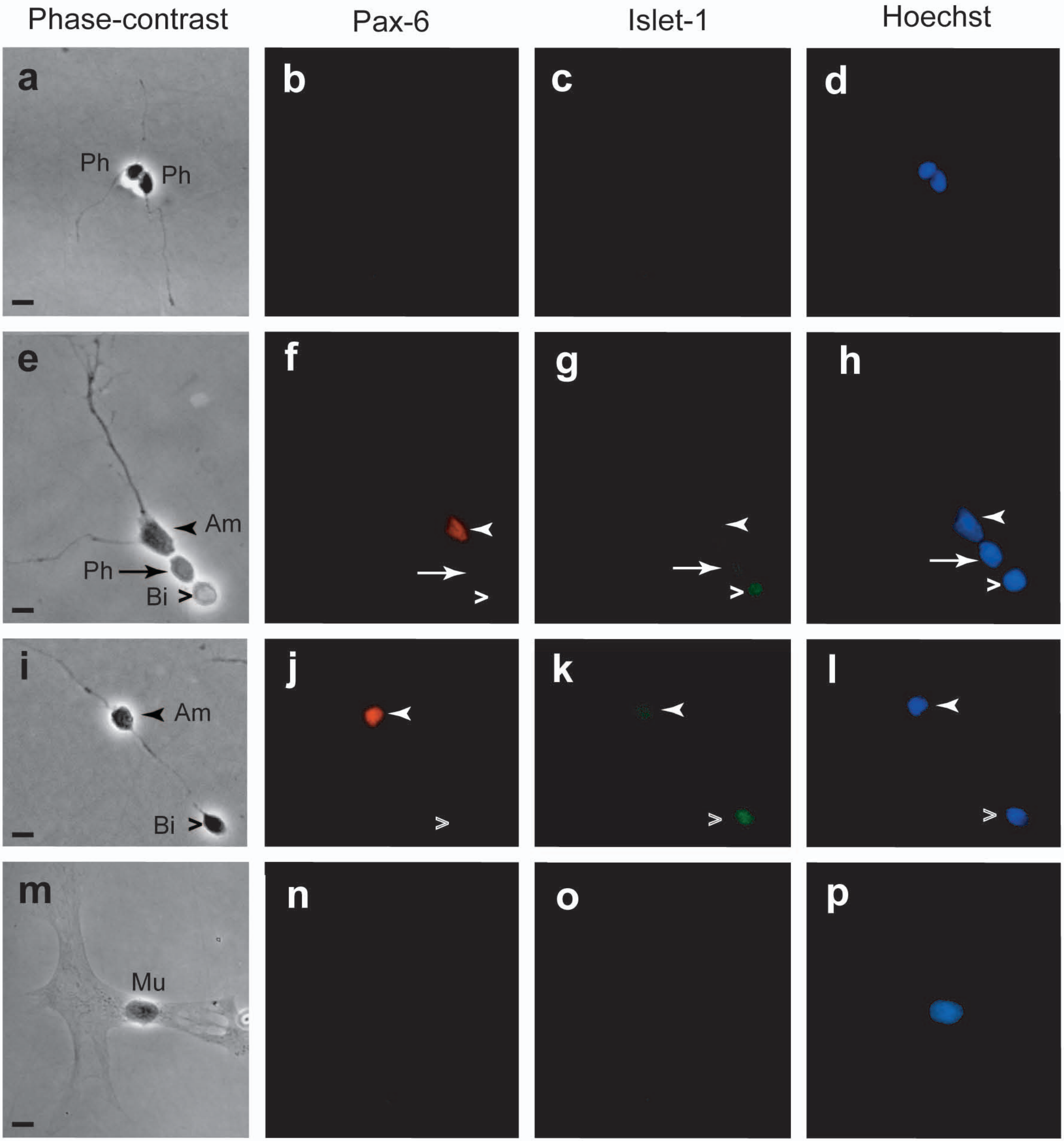
33. M. Cebrian, M. Alfonseca, and A. Ortega, *Information Theory, IEEE Transactions on* **53** (5), 1895 (2007).

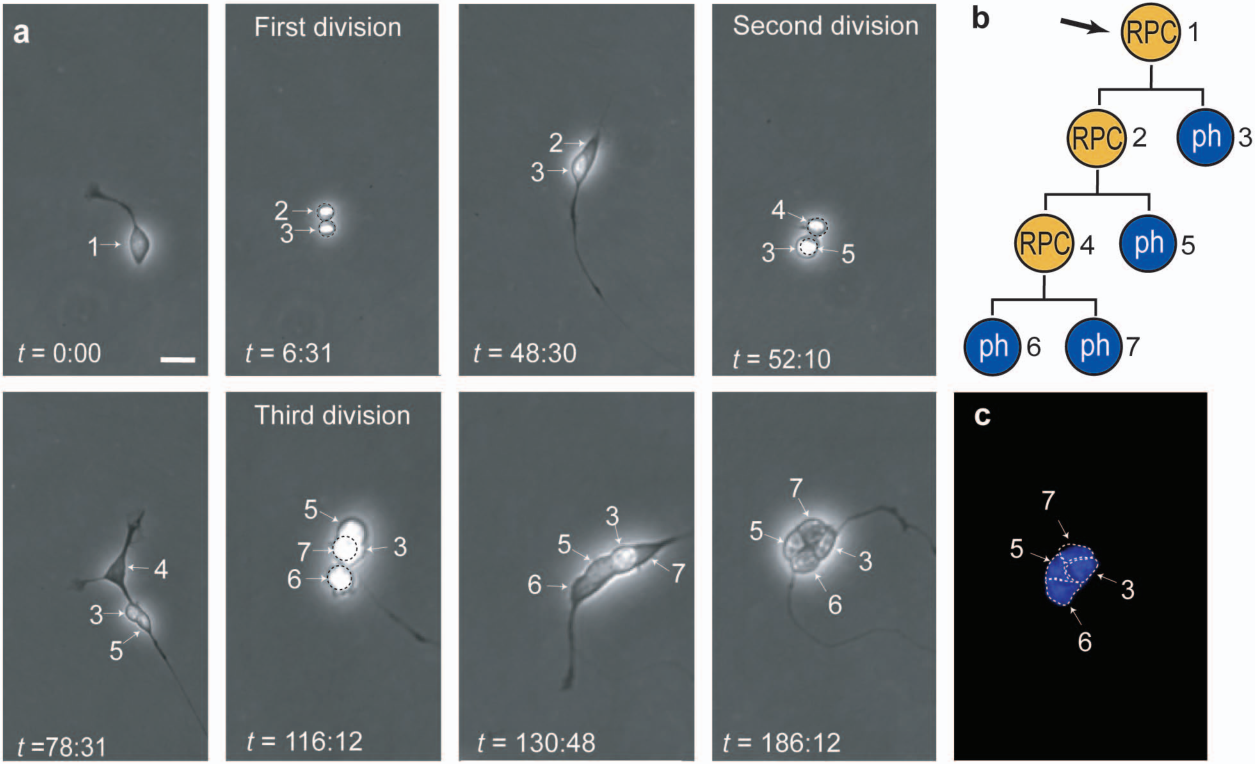
34. Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana, in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM Press, Seattle, WA, USA, 2004).

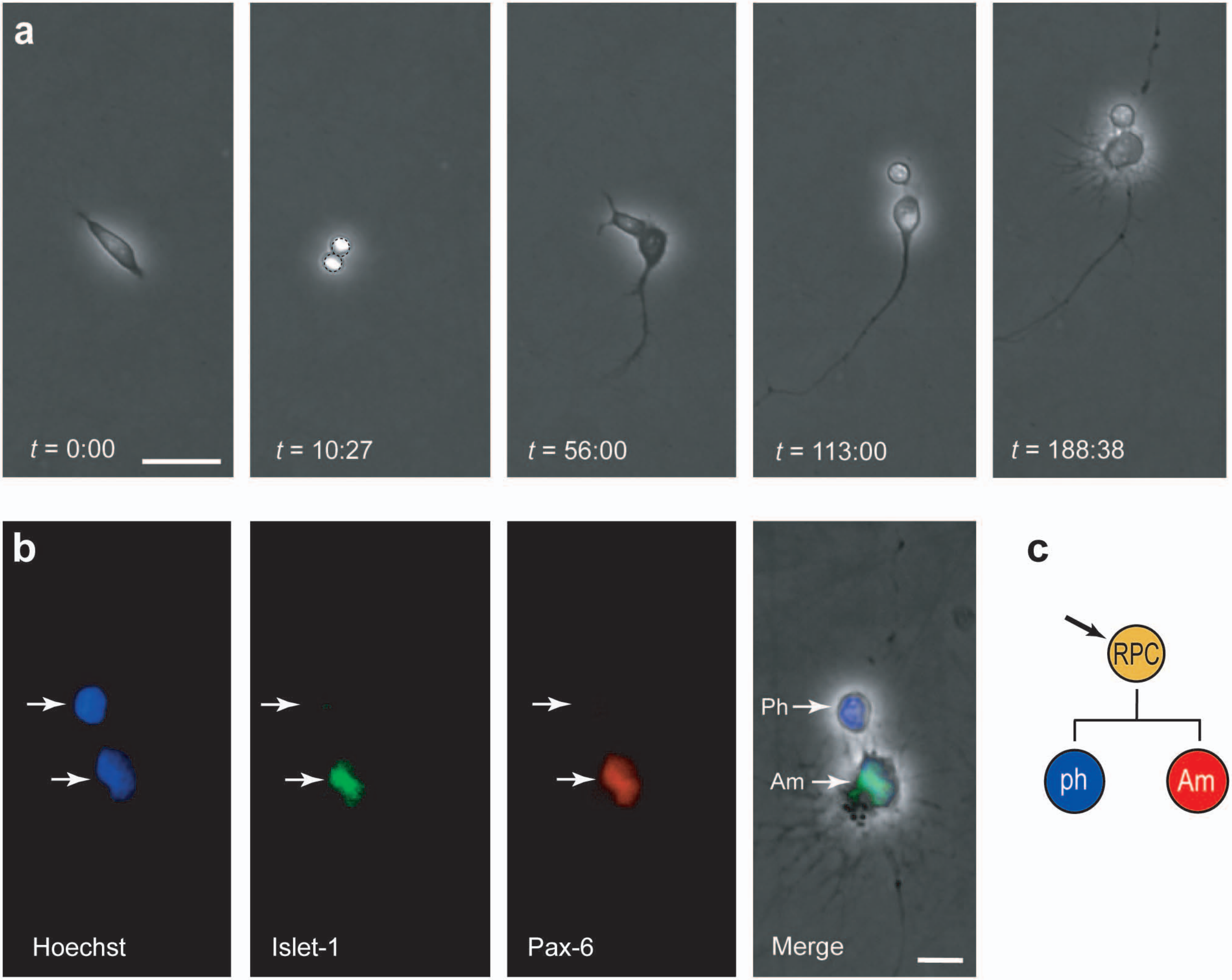
35. J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. (World Scientific, Singapore, 1989).

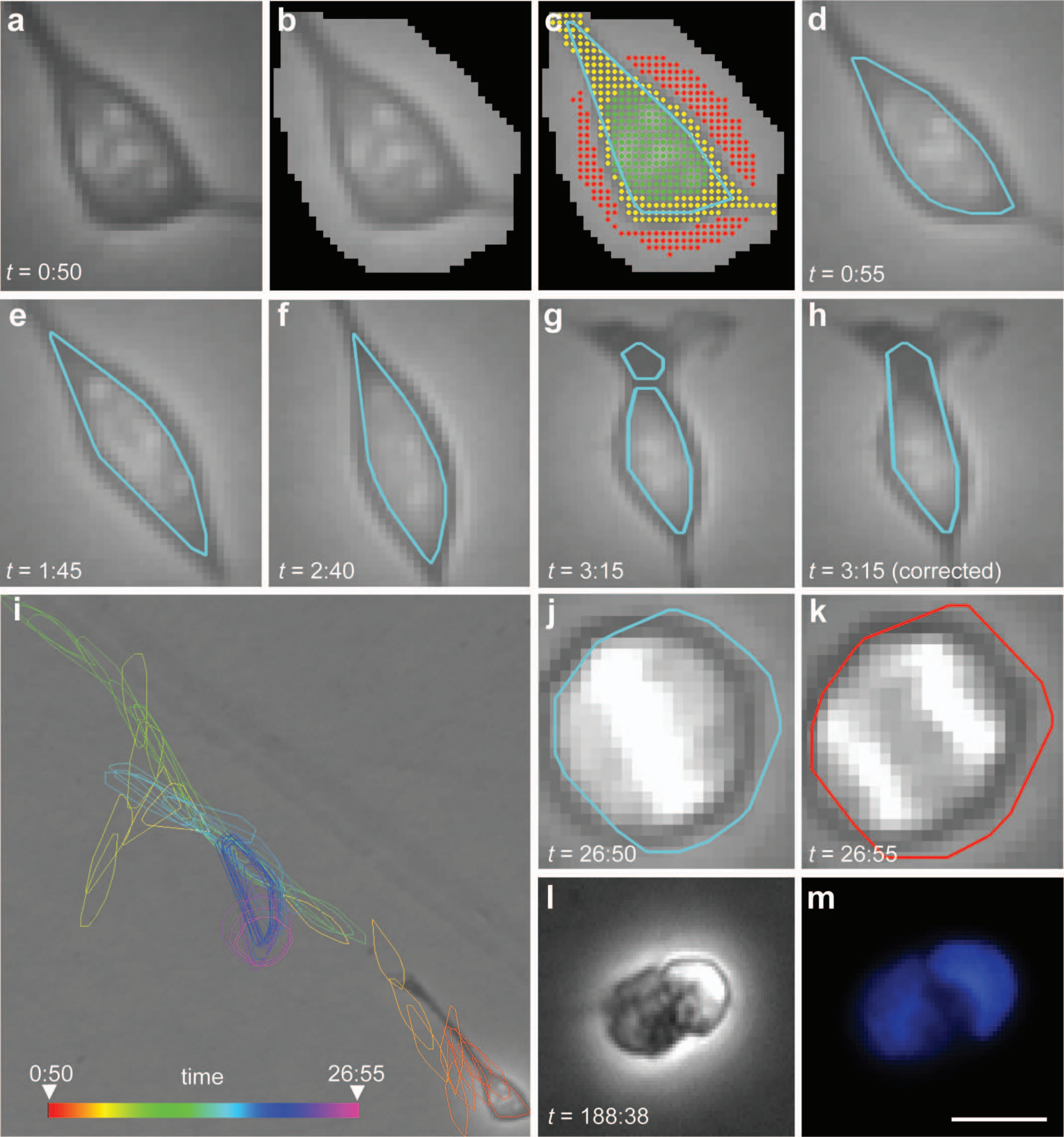
36. Peter Grünwald, In Jae Myung, and Mark Pitt, *Advances in Minimum Description Length: Theory and Applications*. (MIT Press, 2005).

Supplementary File	Title
Supplementary Figure 1	Terminal division of an oligodendrocyte precursor cell (OPC).
Supplementary Figure 2	Self-renewing division of an oligodendrocyte precursor cell (OPC).
Supplementary Table 1	Summary of the criteria used for retinal cell type identification in culture
Supplementary Table 2	The number and type of cells from each of the three experiments included in the prediction of whether an RPC will undergo a self-renewing division or terminally differentiate
Supplementary Table 3	The confusion matrix showing the results of predicting whether an RPC will undergo a self-renewing division or terminally differentiate for Experiment 3, using only Experiments 1 and 2 in the supervised component of the analysis.
Supplementary Video 1	Retial progenitor cell undergoing a self-renewing division (see Fig. 2)
Supplementary Video 2	Retinal progenitor cell undergoing a terminal division (see Fig. 3)
Supplementary Video 3	Example of a retinal progenitor cell segmentation and tracking results
Supplementary Video 4	Oligodendrocyte precursor cell undergoing a terminal division (see Supplementary Figure 1)
Supplementary Video 5	Oligodendrocyte precursor cell undergoing a self-renewing division (see Supplementary Figure 2)

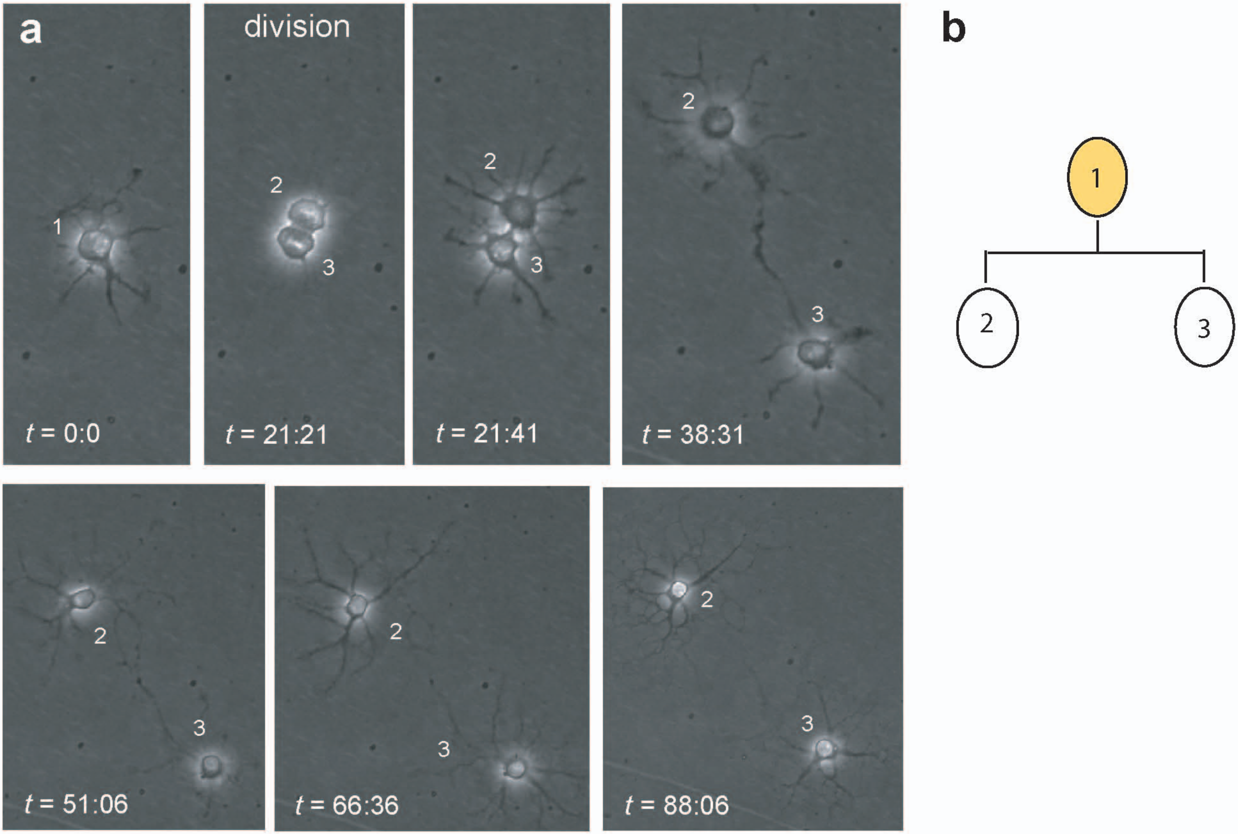

Phase-contrast
Pax-6
Islet-1
Hoechst
a
Ph
Ph
b
c
d
e
Am
Ph
Bi
f
g
h
i
Am
Bi
j
k
l
m
Mu
n
o
p


a
First division
Second division
Third division
t = 0:00
t = 6:31
t = 48:30
t = 52:10
t =78:31
t = 116:12
t = 130:48
t = 186:12
1
2
3
4
5
6
7
b
RPC 1
RPC 2
ph 3
RPC 4
ph 5
ph 6
ph 7
c


a
t = 0:00
t = 10:27
t = 56:00
t = 113:00
t = 188:38
b
Hoechst
Islet-1
Pax-6
Ph
Am
Merge
c
RPC
ph
Am

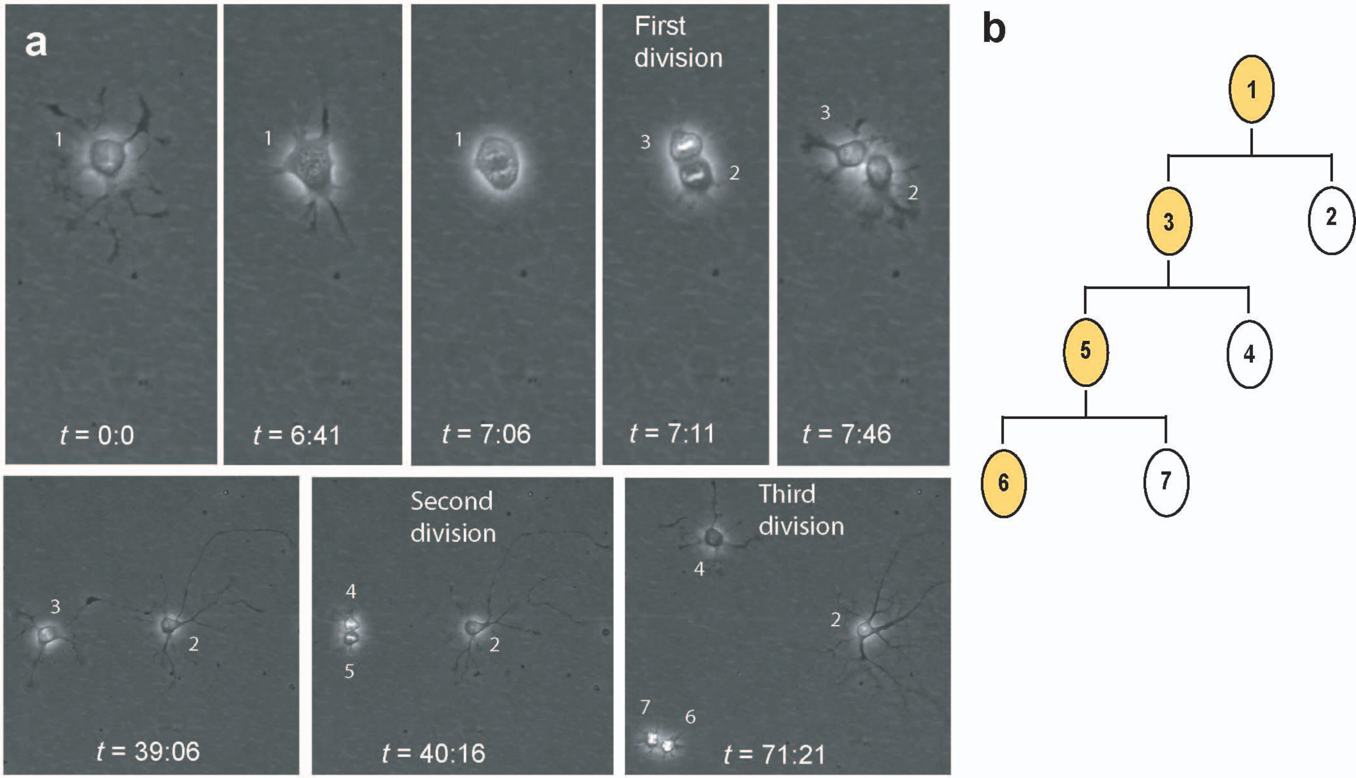

a
t = 0:50
b
c
d
t = 0:55
e
t = 1:45
f
t = 2:40
g
t = 3:15
h
t = 3:15 (corrected)
i
0:50
time
26:55
j
t = 26:50
k
t = 26:55
l
t = 188:38
m

Supplementary Figure 1. Terminal division of an oligodendrocyte precursor cell (OPC).



(a) Snapshots of time-lapse video microscopy showing an OPC giving rise to two oligodendrocytes. (b) Lineage reconstruction of a terminal division. White circles indicate differentiating cells, colored circle indicate proliferating OPC. Time is given in h:min.

Supplementary Figure 2. Self-renewing division of an oligodendrocyte precursor cell (OPC).



(a) Snapshots of time-lapse video microscopy showing the first three divisions. (b) Lineage reconstruction of self-renewing divisions. Cell # 7 will differentiate and cell # 6 will divide again (not shown). White circles indicate differentiating cells, colored circles indicate proliferating OPC. Time is given in h:min.

Supplementary Table 1. Summary of the criteria used for retinal cell type identification in culture

Cell type	Morphology	Islet-1 staining	Pax-6 staining	Condensed chromatin
Photoreceptor	Small, round cell body; one or two thin processes	-	-	+
Amacrine	Large cell body; long branchy processes	+/- * (sub-population)	+	-
Bipolar	Medium-size cell body; one or two thick processes	+	-	-
Müller	Large round nuclei; lamellipodia; lack of neurites	-	- *	-

* Note that a small population of amacrine cells is both Pax6+ and Islet-1+. Since bipolar cells are not Pax-6+, they are easily distinguished from the Pax-6 +/Islet-1+ amacrine cells. Müller cells sometimes stain weakly for Pax-6.

Supplementary Table 2. The number and type of cells from each of the three experiments included in the prediction of whether an RPC will undergo a self-renewing division or terminally differentiate (refer to Table 1).

	Experiment 1	Experiment 2	Experiment 3
Terminal cells	34	0	18
Self renewing cells	7	7	5

Supplementary Table 3. The confusion matrix showing the results of predicting whether an RPC will undergo a self-renewing division or terminally differentiate for Experiment 3, using only Experiments 1 and 2 in the supervised component of the analysis.

		Predicted Outcome	
		Terminal	**Self renewing**
True Outcome	**Terminal**	**18**	**0**
	Self renewing	**1**	**4**

One cell out of 23 was misclassified, giving a prediction accuracy of 96%. Refer to Table 1.