

Computational prediction of neural progenitor cell fates

Andrew R Cohen^{1,2,7}, Francisco L A F Gomes^{3,4,7}, Badrinath Roysam² & Michel Cayouette^{3,5,6}

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 be used to predict cell division outcomes. Here we present a computational method, based on algorithmic information theory, to 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 predicted 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 computer at 5 min per frame. This method could be used to isolate cell populations with specific developmental potential, enabling previously impossible investigations.

In the developing vertebrate nervous system, hundreds of different types of neurons and glial cells are produced from multipotent neural stem and 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 role¹. We and others have proposed that lineage-dependent cell-fate decisions may be 'preprogrammed' 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 heterogeneity in gene expression among mouse RPCs¹⁰, 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 population¹¹, and it is well known that spinal cord development depends on the diversification of the progenitor cell pool¹². Together, these results suggest that a given population of seemingly indistinguishable progenitor cells might actually be composed of many different subpopulations 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 new markers of subpopulations 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.

Here we present a method named algorithmic information theoretic prediction (AITP) that combines recent advances in image sequence summarization¹³ and semisupervised spectral learning¹⁴ to develop this capability. AITP enables sensitive, parameter-free multiresolution analysis of the multidimensional time-sequence data representing the observable spatiotemporal dynamic phenotype of cells obtained from segmentation and tracking algorithms.

¹Department of Electrical Engineering and Computer Science, University of Wisconsin, Milwaukee, Wisconsin, USA. ²Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York, USA. ³Cellular Neurobiology Research Unit, Institut de recherches cliniques de Montréal, Montréal, Canada. ⁴Department of Biology, Universidade Estadual do Ceará, Fortaleza, Ce, Brazil. ⁵Department of Medicine, Université de Montréal, Montréal, Canada. ⁶Departments of Biology and Anatomy and Cell Biology, McGill University, Montréal, Canada. ⁷These authors contributed equally to this work. Correspondence should be addressed to B.R. (roysam@ece.rpi.edu) or M.C. (michel.cayouette@ircm.qc.ca).

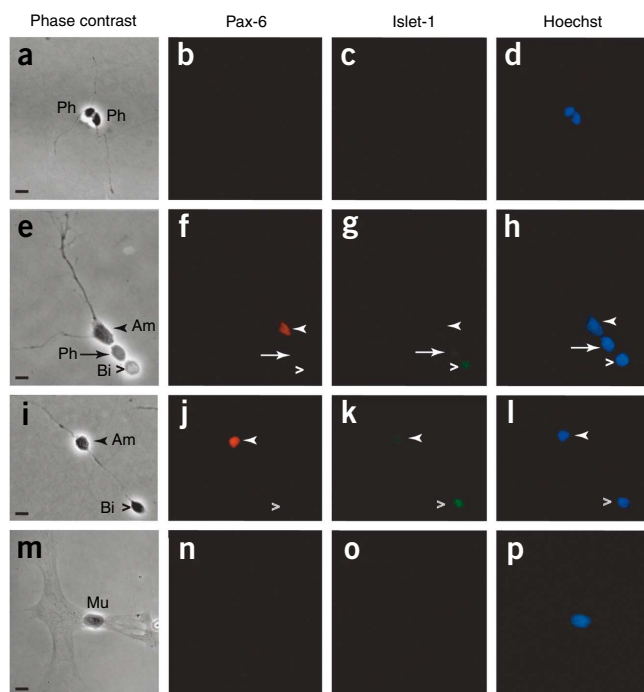


Figure 1 | Retinal cell type identification. (a–p) Phase-contrast and fluorescence images of cells identified based on morphology and by immunostaining for cell type-specific markers Pax-6 and islet-1 as indicated. Hoechst was used to stain DNA. Shown are clones composed of two rod photoreceptors (ph; Pax-6[−] and islet-1[−]) with characteristic round cell body, thin processes and heterochromatin condensation (a–d); one amacrine cell (am; Pax-6⁺ and islet-1[−]), one rod photoreceptor and one bipolar cell (bi; Pax-6[−] and islet-1⁺) (e–h); one amacrine cell and one bipolar cell (i–l); and a Muller glial cell (Mu), with distinct glial morphology, large nucleus, absence of neurites and lack of expression of neuronal markers (m–p). Scale bars, 25 μ m (a–l) and 40 μ m (m–p).

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 and progenitor cells. AITP software is available as **Supplementary Software** and 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 several days. To do this, we grew embryonic day 20 (E20) rat RPCs at clonal density, as previously described². Four hours after plating, we placed the culture dish in a temperature- and CO₂-controlled chamber fitted around an inverted microscope. Then, we identified

presumed RPCs by their characteristic neuroepithelial morphology. Using a motorized microscope stage, we selected >100 cells in each experiment and imaged them by phase-contrast microscopy every 5 min during 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 had previously reported² (**Fig. 1**). We determined the outcome of the first division of each RPC retrospectively by playing back the time-lapse movie and tracking the fate of both daughter cells. Some daughter cells maintained RPC morphology and divided again, indicating that they remained RPCs. Some did not divide again and changed morphology, indicating that they differentiated. In this latter case, we identified retinal cell types as summarized in **Supplementary Table 1**. Overall, we reliably identified the outcome of RPC divisions as either (i) proliferative, generating two RPCs (data not shown); (ii) self-renewing, generating a RPC and a neuron or glial cell (**Fig. 2** and **Supplementary Movie 1**); or (iii) terminally differentiating, generating two neurons of the same or different types (**Fig. 3** and **Supplementary Movie 2**).

Automated measurement of dynamic cell phenotype

We segmented (computationally delineated) the RPCs of interest and tracked them over time (**Fig. 4** and **Supplementary Movie 3**). The results were 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

Figure 2 | Self-renewing divisions.

(a) Snapshots of time-lapse video microscopy showing a rod-committed RPC undergoing two rounds of self-renewing divisions to generate a clone containing four rod photoreceptors; dashed circles depict mitotic cells in telophase. Time is given in hours and minutes. Scale bar, 15 μ m. (b) Lineage reconstruction of the clone shown in a; arrow indicates the progenitor whose behavior was analyzed to generate predictions using AITP. Ph, photoreceptor. (c) Hoechst staining of cell nuclei.

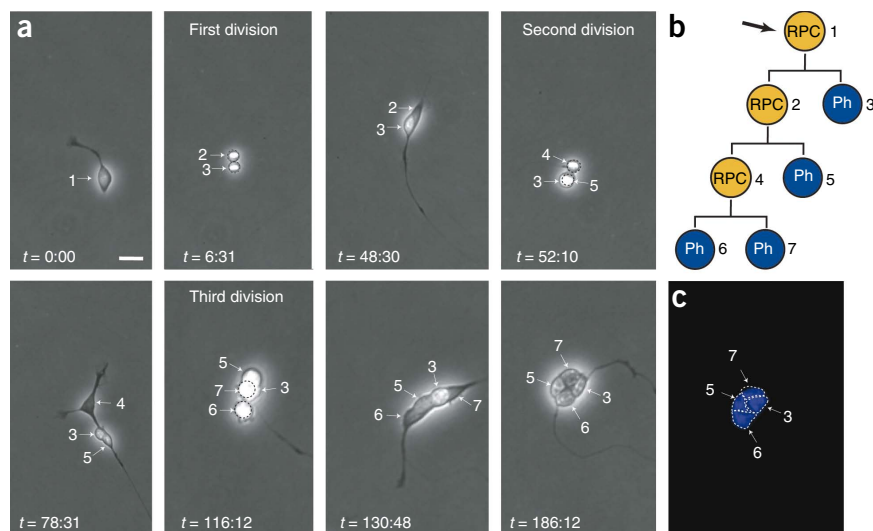
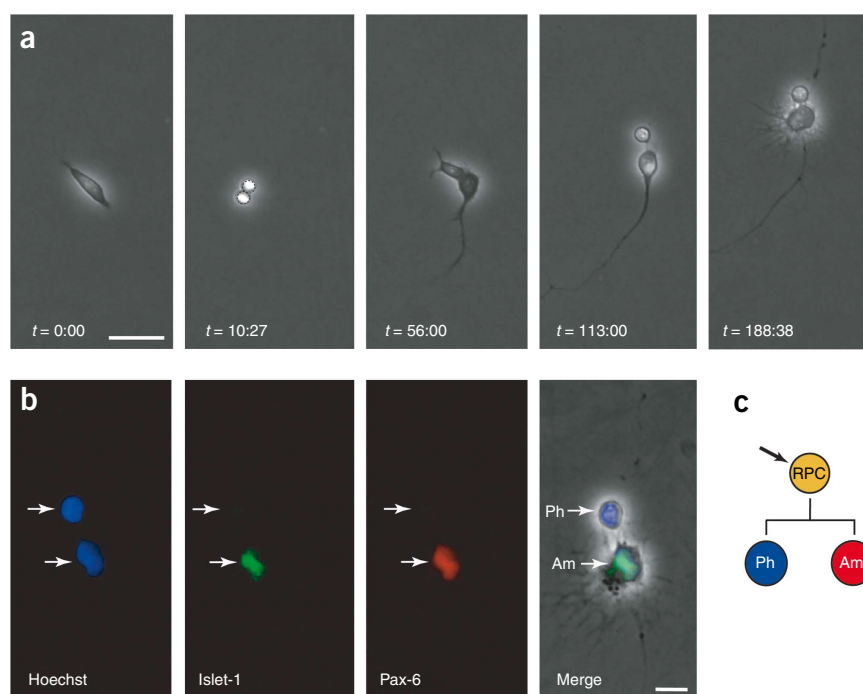


Figure 3 | Terminal division. (a) Snapshots of time-lapse video microscopy showing an RPC undergoing a terminal division to generate a clone containing one rod photoreceptor (ph) and one amacrine cell (am). Dashed circles depict mitotic cells in telophase. Time is given in hours and minutes. Scale bar, 25 μ m. (b) Immunostaining for islet-1 (green) and Pax-6 (red) as well as Hoechst staining (blue). Scale bar, 25 μ m. (c) Lineage reconstruction of the clone shown in a. Arrow points to the RPC whose behavior was analyzed using AITP.



AITP methodology; AITP can be applied directly to any application for which segmentation and tracking algorithms are available. For analyzing RPCs, 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 reported¹⁵. Next, we fit an ellipse to each cell region and compute its eccentricity, $e(t)$. Finally, we compute the size of the cell, $S(t)$, as the area of its convex hull. These calculations yield a time sequence of six 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 (for example, $\theta(t)$, $D(t)$) can improve the accuracy of the compression-based distance measure used in the subsequent analysis. At this stage,

it is unimportant to be concerned with choosing the optimal set of features as irrelevant features are identified and discarded automatically in a later computational step.

Analyzing the dynamic cell phenotype

We compared the dynamic cellular phenotypes, as captured by the time sequences of feature vectors obtained from image segmentation and multitarget tracking algorithms, using the normalized compression distance (NCD) measure. This measure has been reported recently in the field of algorithmic information theory¹⁶, and we enhanced it to include automatic quantization (Online Methods). Briefly, the NCD is used to calculate an $M \times M$ 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 $N = 1 \dots N_{\text{max}}$ 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 $N_{\text{max}} \times (2^{|F|} - 1)$ distance matrices, in which $|F|$ is the number of features. If this number exceeds available computational capacity, feature subset selection methods¹³ can be used to reduce the computational burden.

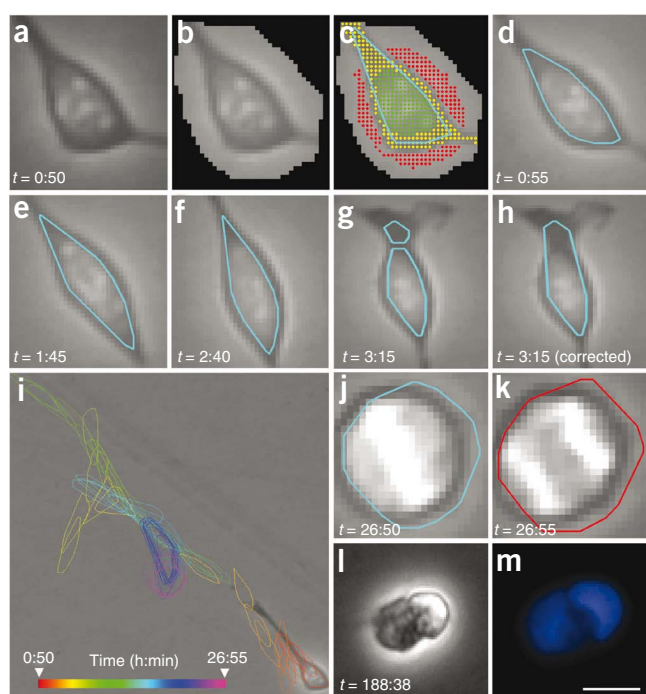


Figure 4 | Automated cell segmentation and tracking. (a) Phase-contrast image of an RPC. (b) The cell is separated from the background based on the halo produced by phase-contrast microscopy. (c) Adaptive clustering separates pixels into groups, with yellow dots indicating dark interior points, green 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 three intermediate times in the time-lapse sequence. (g) Error in segmentation. (h) Automatic correction of the segmentation error from g using the temporal consistency constraint. (i) Overlay of successive color-coded cell outlines superimposed on the initial image (a) to illustrate the cell-tracking results. (j, k) Illustration of mitosis detection. (l, m) Phase-contrast image (l) and Hoechst staining (m) of the progeny after cell division. Scale bar, 7 μ m (a–h); 5 μ m (j, k); 15 μ m (l, m); and 25 μ m (i). The scale bar in m applies to all panels.



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

Confusion matrix					Output parameters	Correct prediction rate	
Self-renewing versus terminal divisions (66 movies, 72 cells)	True outcome	Self-renewing Terminal	Predicted outcome		Quantization = 17 symbols Spectral matrix dimension = 4 Feature subset = (Δx , Δy , θ , S)	99%	
			Self-renewing	Terminal			
			18 ^a 0	1 53			
Photoreceptor (ph) versus non-photoreceptor progeny ^b (76 movies, 86 cells)	True outcome	Non-ph, ph Ph, ph	Predicted outcome		Quantization = 12 symbols Spectral matrix dimension = 80 Feature subset = (Δx , Δy , S)	87%	
			Non-ph, ph	Ph, ph			
			32 9	2 43			
Combinations of progeny: photoreceptor (ph), bipolar (bi) and amacrine (am) ^c (68 movies, 78 cells)	True outcome	Ph, ph Ph, bi Ph, am	Predicted outcome			Quantization = 14 symbols Spectral matrix dimension = 64 Feature subset = (D , S , e)	83%
			Ph, ph	Ph, bi	Ph, am		
			49	0	3		
			4	4	2		
			4	0	12		

^aEntries indicate the number of cell divisions for each category. All three outcomes were analyzed using the same software methodology. ^bResults of predicting whether a terminal RPC will produce two photoreceptor neurons versus another combination of offspring. ^cResults of predicting whether an RPC will produce two photoreceptors versus a photoreceptor-bipolar pair or a photoreceptor-amacrine pair.

This collection of distance matrices is analyzed using a cross-validated semisupervised spectral learning methodology¹⁴ to assign predictions to cells with unknown fate outcomes by comparing their dynamic behaviors to cells from the training data, which have known outcomes (Online Methods). A numeric value is assigned to each possible outcome that we want to predict based on dynamic behavior. For a two-class problem, for which there are two potential outcomes, the assigned value is a 0 or 1; for a three-class problem, where there are three potential outcomes, it is 0, 1 or 2 (Table 1).

Prediction of self-renewing divisions

To determine whether AITP could reliably predict self-renewing versus terminal divisions, we analyzed the time-sequence data of all RPCs for which we could reliably determine the outcome of the first division (Table 1). AITP erroneously predicted only one out of 19 self-renewing RPCs to be a terminal RPC, and it correctly predicted the outcome for all 53 terminal RPCs, corresponding to an accuracy of 99%. The 95% confidence interval for this prediction was 92.5–99.8%¹⁷. We arbitrarily chose a minimum of 100 image frames (about 8 h of recording) for cells to be included in the AITP analysis. The ability to predict whether an RPC will undergo a self-renewing versus a terminal division indicates that RPCs exhibit distinctive dynamic behaviors depending on the type of division they will undergo, and that this pattern can be recognized using AITP.

In the above experiments, we used the entire dataset 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 presented in Table 1 (72 cells) originated from three independent experiments (numbered 1–3 in 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 determined whether AITP could be applied to predict cell fate outcomes of a different neural progenitor cell type. We grew OPCs, which generate oligodendrocytes, the myelinating cells of the central nervous system. These cells, when grown under differentiating conditions (Online Methods), divided to generate two oligodendrocytes (terminal division; Supplementary Fig. 1 and Supplementary Movie 4); one OPC and one oligodendrocyte (self-renewing division; Supplementary Fig. 2 and Supplementary Movie 5); or two OPCs (proliferative division; data not shown). We imaged the cells for 7–10 days, as described for RPCs, and applied AITP to the resulting movies. We predicted whether OPCs would undergo a terminal or a nonterminal 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

As the majority of terminal divisions in the neonatal rat retina produce two photoreceptor cells, we next asked whether we could use AITP to discriminate between terminal divisions generating two photoreceptors or other offspring combinations. 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 1). The 95% confidence interval for this prediction was 78.5–92.7%¹⁷. RPC terminal divisions generated two photoreceptors, a photoreceptor and bipolar neuron or a photoreceptor and amacrine neuron, and AITP could predict the outcome of such divisions with 83% accuracy (Table 1). These results indicate that AITP can be used to predict not only self-renewing versus terminal divisions but also whether a RPC will produce a particular combination of neurons.

Simulated live prediction of cell fate

In the experiments described above, we made the fate predictions retrospectively, after we fixed the cells and identified them by

immunocytochemistry. This was necessary in the first stages of AITP development as it enabled cross-validation of predictions. However, a 'live' fate-prediction functionality for AITP would allow one to isolate cells predicted to have a particular fate during an experiment, before they divide. To add this functionality to AITP, we modified and optimized the algorithm to run at a much faster rate (Online Methods). 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 collecting one image available every 5 min, 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 personal computer (dual quad core Intel Xeon X5472 3-gigahertz processors, 8 gigabytes RAM, running Windows Vista) within the 5-min-per-frame microscope acquisition time. These data required approximately 45 s per frame to segment and update the tracking and an additional 3–5 s 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.

DISCUSSION

The value of our method is its ability to visually identify subtle shape and movement pattern differences. Our results suggest that RPCs in culture display dynamic patterns that can be sensed computationally to predict the outcome of their next division using algorithmic statistics and algorithmic information theory. As AITP could be used to analyze any cell type that can be imaged, segmented and tracked over time, it could potentially have applications in various fields.

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 population¹⁰, these results are difficult to interpret because the fate of any particular RPC was unknown. By applying AITP 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 AITP can also predict the outcome of terminal divisions with high accuracy, it should be possible to isolate RPCs with specific differentiation potential, which might lead to the identification of new genes involved in the specification of various retinal cell types. The cell cycle of neonatal RPCs is ~36–40 h, and we showed that 8 h of recording is sufficient to generate an accurate prediction. As the majority of RPCs would not undergo mitosis in the first 8 h 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 two-dimensional, and even three-dimensional and multichannel time-lapse data¹⁸. As proof-of-principle that AITP can be extended to other cell types, we used AITP to accurately predict cell division pattern of OPCs. 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 in the feature subset that differentiate the populations of interest is a topic for future research. From an algorithmic information theoretic standpoint, we estimate 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 identifies 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, we wrote our software to take advantage of any parallel computing facility that supports the common message passage interface (MPI) library. We also provide a version that runs on serial computers for generality. We 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.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

The computational aspects of this work were supported by the Center for Subsurface Sensing and Imaging Systems (US National Science Foundation; EEC-9986821), by the Rensselaer Polytechnic Institute and by the University of Wisconsin–Milwaukee. We thank A. Todorski and the staff at Rensselaer's Computational Center for Nanotechnology Innovations supercomputing center; S. Temple for helpful comments and feedback, and on collaborations leading up to the present work; M. Kmita and J. Chan for insightful comments on the manuscript; A. Daigneault for expert technical assistance with the animal colony; and C. Jolicœur 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 supported by the Canadian Institutes of Health Research New Investigator program and the W.K. Stell scholarship from the Foundation Fighting Blindness, Canada.

AUTHOR CONTRIBUTIONS

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

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

1. Cayouette, M., Poggi, L. & Harris, W.A. Lineage in the vertebrate retina. *Trends Neurosci.* **29**, 563–570 (2006).
2. Cayouette, M., Barres, B.A. & Raff, M. Importance of intrinsic mechanisms in cell fate decisions in the developing rat retina. *Neuron* **40**, 897–904 (2003).
3. Godinho, L. *et al.* Nonapical symmetric divisions underlie horizontal cell layer formation in the developing retina in vivo. *Neuron* **56**, 597–603 (2007).
4. Mu, X. *et al.* Ganglion cells are required for normal progenitor-cell proliferation but not cell-fate determination or patterning in the developing mouse retina. *Curr. Biol.* **15**, 525–530 (2005).
5. Poggi, L., Vitorino, M., Masai, I. & Harris, W.A. Influences on neural lineage and mode of division in the zebrafish retina in vivo. *J. Cell Biol.* **171**, 991–999 (2005).
6. Diaz, E. *et al.* Analysis of gene expression in the developing mouse retina. *Proc. Natl. Acad. Sci. USA* **100**, 5491–5496 (2003).
7. Dorrell, M.I., Aguilar, E., Weber, C. & Friedlander, M. Global gene expression analysis of the developing postnatal mouse retina. *Invest. Ophthalmol. Vis. Sci.* **45**, 1009–1019 (2004).
8. Livesey, F.J., Young, T.L. & Cepko, C.L. An analysis of the gene expression program of mammalian neural progenitor cells. *Proc. Natl. Acad. Sci. USA* **101**, 1374–1379 (2004).
9. Mu, X. *et al.* Gene expression in the developing mouse retina by EST sequencing and microarray analysis. *Nucleic Acids Res.* **29**, 4983–4993 (2001).
10. Trimarchi, J.M., Stadler, M.B. & Cepko, C.L. Individual retinal progenitor cells display extensive heterogeneity of gene expression. *PLoS One* **3**, e1588 (2008).
11. Tietjen, I. *et al.* Single-cell transcriptional analysis of neuronal progenitors. *Neuron* **38**, 161–175 (2003).
12. Jessell, T.M. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nat. Rev. Genet.* **1**, 20–29 (2000).
13. Cohen, A.R., Bjornsson, C.S., Temple, S., Banker, G. & Roysam, B. Automatic summarization of changes in biological image sequences using algorithmic information theory. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 1386–1403 (2009).
14. Kamvar, S.D., Klein, D. & Manning, C.D. Spectral learning. *International Joint Conference of Artificial Intelligence* (2003).
15. Baye, L.M. & Link, B. Interkinetic nuclear migration and the selection of neurogenic cell divisions during vertebrate retinogenesis. *J. Neurosci.* **27**, 10143–10152 (2007).
16. Cilibrasi, R. & Vitanyi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Theory* **51**, 1523–1545 (2005).
17. Witten, I.H. & Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, 2005).
18. Chen, Y., Ladi, E., Herzmark, P., Robey, E. & Roysam, B. Automated 5-D analysis of cell migration and interaction in the thymic cortex from time-lapse sequences of 3-D multi-channel multi-photon images. *J. Immunol. Methods* **340**, 65–80 (2009).



ONLINE 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 Montréal animal care committee. Retinal cells from embryonic day 20 (E20) Sprague Dawley rat retinas were cultured as previously described². Approximately 2×10^4 cells were plated in 35 mm Falcon dishes coated with poly(L-lysine) ($10 \mu\text{g ml}^{-1}$) and laminin ($10 \mu\text{g ml}^{-1}$). The dissociated cells were allowed to settle for 4 h in a CO_2 incubator at 37°C before they were placed under the time-lapse microscope. RPCs were imaged with a Leica $20\times$ phase-contrast objective, and the images captured using a Hamamatsu charge-coupled device (CCD) video camera connected to a Macintosh computer equipped with Volocity software (Improvision) programmed to capture a frame every 5 min. The cells were kept at 37°C , 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 d, and the following antibodies were used for immunofluorescence: monoclonal mouse anti-islet-1 (1:2,000; produced by T. Jessell and obtained from the Developmental Studies Hybridoma Bank) and rabbit anti-Pax-6 (1:10,000; 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 20 optic nerves were removed from postnatal day 7 (P7) Sprague Dawley rats and dissociated and cultured as described previously¹⁹. Before time-lapse imaging, half of the medium was replaced with MBS medium supplemented with T3 (thyroid hormone; Sigma; 40 ng ml^{-1}) to induce OPC differentiation²⁰. 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 imaged in phase contrast exhibit a characteristic halo that we exploited for segmentation. Pixels whose brightness was 5 s.d. above the average pixel intensity were considered halos or bright interior pixels. Morphological opening and closing operations²¹ were 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 were then separated into connected regions using the watershed transform²². The morphological smoothing prevents unwanted fragmentation²³. A spectral k -means clustering²⁴ 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 and the intermediate-intensity halo pixels (Fig. 4c) were discarded. A constrained watershed transform was applied to the cell pixels to separate touching cells. The watershed transform was constrained using the H-maxima transform²¹ to reduce fragmentation.

To track cells, we used a variation of the method described previously²⁵. We used tracking anomalies to identify scenarios in

which the watershed transform has resulted in erroneous cell fragmentation and then merged the oversegmented regions automatically. This was implemented using a temporal look-ahead step whenever the assignment matrix used by the tracking algorithm was not square. In that case, our algorithm checks for temporal consistency up to 9 frames ahead in the video sequence to see whether 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 images as bright regions of approximately equal size. Using morphological opening and connected component analysis, we examined whether there were two such bright regions (Fig. 4j,k). The first time that two such regions are detected is considered the time of mitosis (Fig. 4k). Before applying AITP, the results were screened visually for tracking errors. Cells with tracking errors were excluded from analysis.

There are many 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 theoretically ideal distance measure for comparing arbitrary datasets is the normalized information distance³⁰ based on Kolmogorov's algorithmic information theory³¹ that can account for any and all differences. However, this measure is impractical owing to its uncomputability³². To overcome this limitation, the normalized compression distance (NCD) had been proposed as a practical approximation to the normalized information distance using data compression algorithms¹⁶. 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,

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

The NCD is unique in its ability to accurately compare multi-dimensional 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 has been shown that the choice of compression algorithm has a negligible impact on the final analysis¹⁶. 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 automatically¹³. This procedure enables a multiresolution analysis because 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 in **Supplementary Software**.

Method to transform the NCD distance matrices to eigenspace.

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

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

The normalized matrix $\mathbf{L}$ is symmetrical with zeros on the diagonal (that is, it is positive semidefinite). The eigenvectors of $\mathbf{L}$ are sorted by the magnitudes of the eigenvalues and used to form the columns of a new $M \times M$ matrix denoted $\mathbf{V}$. 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 multidimensional time-sequence data in a form that naturally enables the use of unsupervised and/or supervised machine learning algorithms (clustering, nearest neighbors, decision trees and others).

Semisupervised analysis of the $\mathbf{L}$ matrix with cross-validation.

Unsupervised analysis methods, as previously described¹³, are appropriate for exploring unlabeled data, that is, 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 semisupervised 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 $\mathbf{L}$ 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 learning¹⁴ methods. This is the method we adopted.

We used the following semisupervised 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 two-class problem this is a 0 or 1; for a three-class problem, it is 0, 1 or 2 (**Table 1**).

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-validation¹⁷. The validation error for classifier C^j is estimated as:

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

Using the above procedure, we identify the values of f , k , N and C that minimize the validation error.

Software source code written in C (both single processor and MPI versions) for performing the above analysis is available as **Supplementary Software**. More details about the software are available 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 postmitotic outcomes. Before applying live-cell-fate prediction, AITP was 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.

- Barres, B.A. *et al.* Cell death and control of cell survival in the oligodendrocyte lineage. *Cell* **70**, 31–46 (1992).
- Barres, B.A., Lazar, M.A. & Raff, M.C. A novel role for thyroid hormone, glucocorticoids and retinoic acid in timing oligodendrocyte development. *Development* **120**, 1097–1108 (1994).
- Soille, P. *Morphological Image Analysis: Principles and Applications* (Springer-Verlag, 1999).
- Vincent, L. & Soille, P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 583–598 (1991).
- Lin, J., Keogh, E., Lonardi, S. & Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. *Data Min. Knowl. Discov.* **15**, 107–144 (2007).
- Ng, A.Y., Jordan, M. & Weiss, Y. On Spectral Clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst* **14**, 849–856 (2001).
- Al-Kofahi, O. *et al.* Automated cell lineage tracing: a high-throughput method to analyze cell proliferative behavior developed using mouse neural stem cells. *Cell Cycle* **5**, 327–335 (2006).
- Debeir, O., Van Ham, P., Kiss, R. & Decaestecker, C. Tracking of migrating cells under phase-contrast video microscopy with combined mean-shift processes. *IEEE Trans. Med. Imaging* **24**, 697–711 (2005).
- Jaqaman, K. *et al.* Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods* **5**, 695–702 (2008).
- Li, K. *et al.* Cell population tracking and lineage construction with spatiotemporal context. *Med. Image Anal.* **12**, 546–566 (2008).
- Meijering, E., Smal, I. & Danuser, G. Tracking in molecular bioimaging. *IEEE Signal Process. Mag.* **23**, 46–53 (2006).
- Bennett, C.H., Gacs, P., Ming, L., Vitanyi, M.B. & Zurek, W.H. Information distance. *IEEE Trans. Inf. Theory* **44**, 1407–1423 (1998).

31. Li, M. & Vitanyi, P.M.B. *An Introduction to Kolmogorov Complexity and Its Applications* (Springer Verlag, New York, 1997).
32. Li, M., Chen, X., Li, X., Ma, B. & Vitanyi, P.M.B. The similarity metric. *IEEE Trans. Inf. Theory* **50**, 3250–3264 (2004).
33. Cebrian, M., Alfonseca, M. & Ortega, A. The normalized compression distance is resistant to noise. *IEEE Trans. Inf. Theory* **53**, 1895–1900 (2007).
34. Keogh, E., Lonardi, S. & Ratanamahatana, C.A. Towards parameter-free data mining. in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, Seattle, 2004).
35. Rissanen, J. *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989).
36. Grünwald, P., Myung, I.J. & Pitt, M. *Advances in Minimum Description Length: Theory and Applications* (MIT Press, 2005).