C. ELEGANS CELL MATCHING AND TRACKING IN A 4D IMAGEING SYSTEM

Long Chen[1]*, Zhongying Zhao*[2] *and Hong Yan*[1]
[1]Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
[2]Department of Biology, Hong Kong Baptist University, Kowloon, Hong Kong

ABSTRACT

Automatic cell tracking for time-lapse images becomes more and more important for live cell studies because the manual tracking is extremely time consuming. In this paper, we proposed a method for *C.elegans* cell tracking based on probabilistic relaxation labeling (PRL). The experiment results obtained in this research indicate that our method could track the *C.elegans* cells with a high accuracy at a very low time resolution. Our method provides an efficient tool for the analysis of high-throughput large *C. elegans* microscopy image data sets.

Index Terms— Cell tracking, *C. elegans*, probabilistic relaxation labeling.

1. INTRODUCTION

Three-dimensional (3D) time-lapse images of *C. elegans* embryos provide highly valuable information for functional genomics studies [1-3]. According to these 3D images, a cell lineage tree of *C. elegans* could be built, which could be used for further biological analysis at a single cell resolution. However, due to the massive amounts of imaging data, it is almost impossible to build the lineage tree manually. So an efficient algorithm for automatically cell tracing is needed for the lineage tree generation. In this paper, we propose a method based on the probabilistic relaxation labeling (PRL) to track the cells.

The greatest challenge for cell lineage tracking comes from the fact that at the end stage of the embryogenesis, the cell number is over 500 and almost all of these cells are crowed together. In addition, in order to get more imaging dataset, one microscope will collect images from 3 to 4 different embryos simultaneously. Therefore, the time resolution (60 – 90 second per time point) is not enough for a precise definition of all the one-to-one or one-to-two (division) matching since even in two adjacent time points, the same cell may move a long distance or divided suddenly. Our method focuses on these two problems, which tries to do cell tracking with a high accuracy rates at very late cell stage (>500) and very low time resolution (>180s).

The most popular *C. elegans* cell lineage tracking tool is StarryNite [4] and its later version StarryNite II [5]. StarryNite and StarryNite II are all based on nearest neighbor (NN) matching, and StarryNite II introduces a layered greedy approach to correct the tracking errors. Another method is based on Support Vector Machines. In this tracking method, many other features like cell radius also have been used to improve the tracking performance [6]. Other method with multiple active surfaces [10] and model-based approach [11] only perform well at early stage (<180). When more cells become crowded together, the tracking errors will explode.

In this paper, we propose an algorithm for cell tracking in the *C. elegans* 4D imaging system based on the PRL. In our method, firstly we transfer the tracking problem into a one-to-one non-rigid point matching problem. Then we used the relative position information to represent each cell. The reason why relative position is better than absolute position is that the relative position could represent the local structure of embryo. The local structure of the embryo will not change a lot within a short time. The experiment results prove that our method based on the relative position could tracking cells with a low time resolution with a high accuracy. Furthermore, our method also has a low computational complexity.

This paper is organized as follows. In Section 2, we explain how to treat the cell tracking problem as a non-rigid point matching problem. Section 3 describes the PRL point matching framework used to tracking the cell in *C. elegans* data. Section 4 presents the experiment results. Section 5 provides conclusions and directions for future work.

2. PROBLEM FORMULATION

A cell may appear at adjacent time points or divide into two new cells during the time interval. Before cell tracking, the original 3D time-lapse image will be processed by a 3D segmentation process after which the 3D position of every cell at every time point will be obtained for the cell lineage tracking [9]. Then the cell-tracking procedure tries to define the relationship of cells at different time points by tracing the cell at one by one time point and finally produces a cell lineage tree, which contains a lot of valuable information such as the cell cycle, gene expression value and 3D position etc., which could be used for further biological analysis [4]. However, at the late stage of embryogenesis, hundreds of cells will be crowded together. So when the cell move a distance longer than the cell diameter, which easily

happens when the time resolution is low, the nearest neighbor (NN) matching will make many tracking errors.

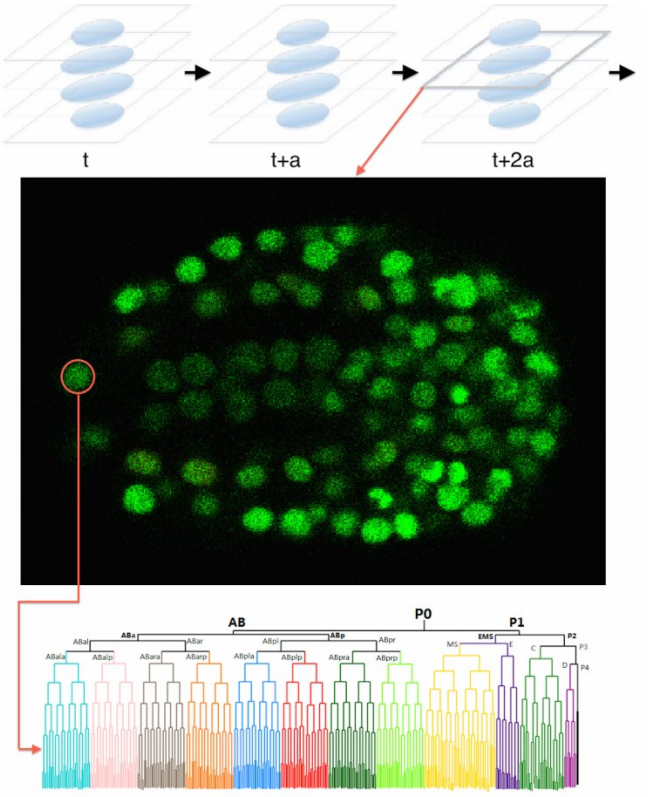

Figure 1. A raw cell image and the cell lineage tree.

In our method, we convert the cell-tracking problem into a point-matching problem. Let us define the cells detected in time point t as a set of points $Q_t = \{q_1, q_1, \dots, q_n\}$, and the cells detected in the next time point $t+1$ as another set of points $S_{t+1} = \{s_1, s_2, \dots, s_m\}$. The number of cells at time point $t+1$ should be equal or larger than that at time point t because of cell division. Therefore, additional points are introduced in time point t to represent the "dummy cells". Then the first point sets will be $Q'_t = \{q_1, q_1, \dots, q_n, q'_{dummy}\}$. The new divided cells in time point $t+1$ could be matched to dummy cell q'_{dummy} and other non-divided cell can be matched one-to-one. Thus, the tracking problem becomes a non-rigid point-matching problem, the task is to find out all the one-to-one matching between every point sets.

In order to solve the non-rigid point-matching problem, a method employing probabilistic relaxation labeling (PRL) is adopted in this work [7-8]. Based on the relative position, the matching probability between q_i and s_j will be initialized according to their Euler distance with other cells at the same time point. Compared with absolute position, the relative distance is a better feature than for tracking which could characterize the local structure of the embryo. In other words, the local structure matching is better than the nearest neighbor (NN) matching. It is because that, the cell could move a long distance, but cells will never exchange their position with each other during a short time period. As illustrated in figure 2, cell q_z at time point t would wrongly match to cell s_y at time point $t+1$ based on the nearest neighbor (NN) matching method. However, the relative position shows that the cell s_y should match to cell s_x at time point t. So our method based on the relative position would have a better performance when tracking cells at a low time resolution.

3. PROPOSED METHOD

The cell-tracking procedure using our PRL-based tracking approach contains 3 steps:

(1) Computing the compatibility coefficient based on cell-to-cell distance;
(2) Matching probability initialization;
(3) Relaxation labeling iteration.

3.1 Computing the compatibility coefficient

As discussed before, the relative position information is used for the initialization of the matching process. So firstly at every time point, we calculate all the Euclidean distance between every point pair. As shown in Figure 2, cell q_i and q_x are at time point t and cell s_j and s_y at time point $t+1$, a compatibility coefficient can be determined as

$$c_{ijxy}^{t,t+1} = 1 - \left| \frac{d_t(q_i, q_x) - d_{t+1}(s_j, s_y)}{d_t(q_i, q_x) + d_{t+1}(s_j, s_y)} \right|, \quad (1)$$

where $d_t(q_i, q_x)$ is the distance between cell q_i and q_x, $d_{t+1}(s_j, s_y)$ is the distance between cell s_j and s_y. The compatibility of cell q_i matching cell s_j and cell q_x matching cell s_y is represented by the compatibility coefficient $c_{ijxy}^{t,t+1}$ ranging from 0 to 1. A high value of $c_{ijxy}^{t,t+1}$ means a high compatibility between q_i matching s_j and q_x matching to s_y.

3.2 Matching probability initialization

The initialization is essential in our method. It is because that the probabilistic relaxation labeling (PRL) method will only converge to a local optimal matching solution based on the initialization result. According to the compatibility coefficient defined above, the probability of cell q_i matching cell s_j is given by

$$p_{ij} = \sum_{x=1}^{n} \sum_{y=1}^{m} c_{ijxy}^{t,t+1}, \quad (2)$$

Note that p_{ij} is supported by $c_{ijxy}^{t,t+1}$, which means that the probability of cell q_i matching cell s_j depends on whether the matching is compatible with other matching pairs.

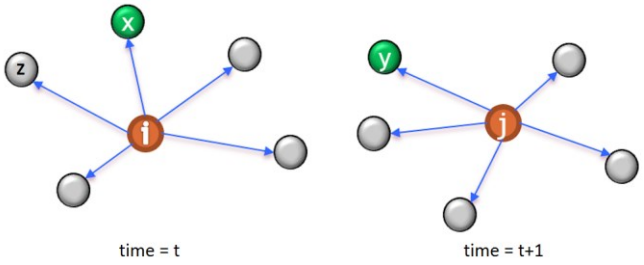

Figure 2. Matching cell pairs at two time points.

Finally the overall correspondence matrix between time point t and time point $t + a$ is determined as fellow:

$$P_{t,t+1} = \begin{bmatrix} p_{11} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nm} \end{bmatrix}, \quad (3)$$

Each row will be normalized before the third step so that $p_{ij} \in [0,1]$.

3.3 Relaxation labeling interaction

An initial correspondence matrix is defined after the above two steps. However, correspond matrix $P_{t,t+1}$ only represents a soft matching relationship, since the matching probability p_{ij} in this matching matrix is ranged from 0 to 1. In order to obtain a certain matching relationship, the matching matrix $P_{t,t+1}$ will iteratively update according to the following equation. The iteration is carried out as follows and will stop if all the elements in the matching matrix $P_{t,t+1}$ are close to 0 or 1:

$$p_{ij}^{k+1} = \frac{\sum_{x=1}^{n}\sum_{y=1}^{m} c_{ijxy}^{t,t+1} p_{ij}^{k}}{\sum_{x=1}^{n}\sum_{y=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ijxy}^{t,t+1} p_{ij}^{k}}, \quad (4)$$

where p_{ij}^{k+1} is the probability of cell q_i matching cell s_j after (k+1)-th iteration. According to the update equation, the probability of each correspondence relationship between two cells is dependent on the other cell pairs. If a corresponding relationship is supported by other correspondence cell pairs, the probability of this matching relationship will increase. Otherwise the probability will decrease. Finally, when the matching matrix only contains values close to 0 or 1, a one-to-one matching result is obtained [7,13].

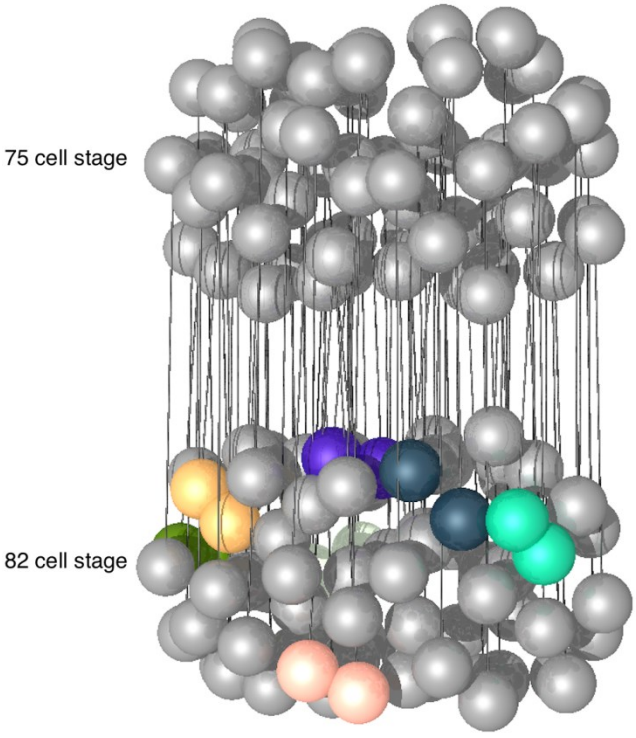

Figure 3. 3D representation of cell matching between two adjacent cell stages. New cells from cell division are shown in color.

Due to the cell division, sometimes the adjacent two time points may have different numbers of cells. So if the cell numbers of two adjacent time point are not equal, we will add dummy cell in the first point set, as discussed in Section 2. For example, if cell s_j at time point $t + 1$ does not match to a real cell according to the correspondence matrix $P_{t,t+1}$, which means that this cell matches to a dummy cell, cell s_j will be identified as a new cell generated from cell division. Based on this kind of idea, all these cell divisions could be detected.

4. EXPERIMENTAL RESULT

The efficiency of our method was evaluated with 2 embryo imaging data sets, each containing 200 time points with a time resolution of 90 seconds. The number of tracking required for the built up of one lineage tree is over 20,000 and a total of more than 50,000 tracking of cell pairs are performed in this analysis for two embryo imaging data sets. As discussed before, cell segmentation has been performed before cell tracking to determine all the cell positions at every time points [9]. The cells position data were further checked manually to ensure that all the cell position data are error-free.

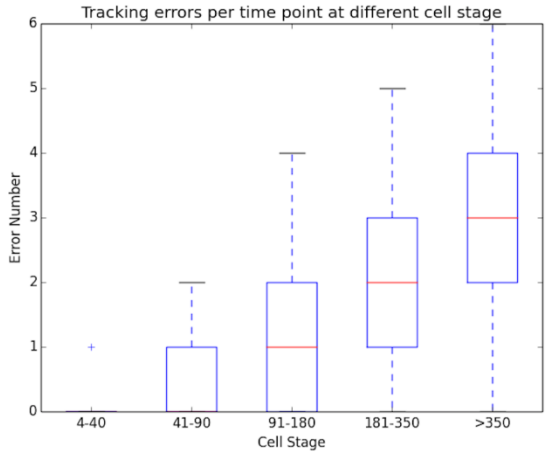

Figure 4. The number of tracking errors per time point at different cell stages.

A high accuracy rate is extremely important because even 1% error rate will cost hours of time for manfully editing the lineage. For example, a laboratory scientist needs 10 to 30 seconds to identify and correct one tracking error caused by the tracking algorithm, which means that one image data set may need $20{,}000 \times 1\% \times 10 \approx 3$ hours to editing before further biological studies. The accuracy of our method is analyzed by comparison of our tracking solution and the error-free manually built lineage trees. The comparison result is shown in Figure 4. At < 350 cell stage, our tracking accuracy rate is over 99.7%. Despite the decreasing tendency of accuracy rate as the number of cells at one time point increases, our method can still maintain a high accuracy rate even at very late stage. As demonstrated in Figure 4, there is an average of only 3 tracking errors at > 350 cell stage.

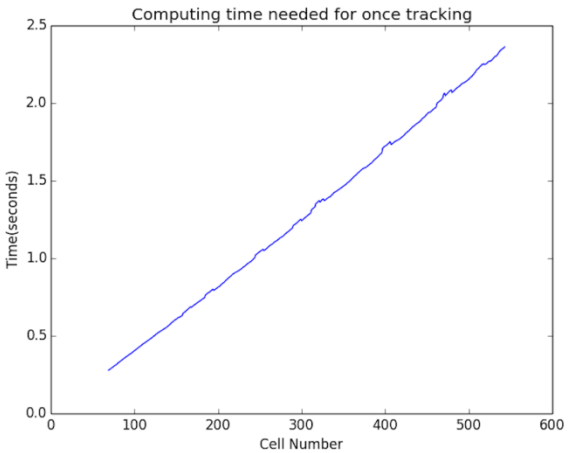

Figure 5. Computing time for one tracking at different cell stages.

Another important performance indicator for a tracking method is the computational complexity, which is also verified in our experiment. Computational complexity is important because of the large number of data sets. A low computational complexity is really needed when for high-throughput analysis. In this paper, our method is implemented in Python and C++. In Figure 5, the time required for one tracking is presented, which is largely determined by the number of cells at a time point. On a 3.47 GHz PC with 10GB of RAM, the computing time of our method for once matching is 0.3 to 2.3 seconds with a single threaded process. For one full lineage tracking from 4 to 550 cell stages, the overall required time is about 720 seconds, which is quick enough for a high-throughput analysis. Furthermore, because the matching processes are independent, it is easy to realize our method with multi-core CPUs (Table 1) or GPUs. So we believe that our method is also suitable for high-throughput image data analysis.

Table 1 Multi-Core test

	1 core	4 cores	12 cores
Time (seconds)	720.4	217.8	85.9

5. CONCLUSION

In this paper, we have introduced an automatic cell tracking method based on probabilistic relaxation labeling (PRL). The cell tracking problem was treated as a point-point matching problem. The experiment results with the *C. elegans* image data show that our method could reach an average of 99% accurate rate at a low time resolution (90 seconds). Furthermore, our method has a high speed, making it suitable for parallel computing, which can provide a significant advantage for the high-throughput image analysis. Future research should focus not only on the using the probabilistic relaxation labeling (PRL) for other type of image data like zebra-fish and plant, but also comparing the cell embryo structures.

Acknowledgement: This work is supported by the Hong Kong Research Grants Council (Project HKBU5/CRF/11G) and City University of Hong Kong (Project 9610308).

6. REFERENCES

[1] Brenner S. The genetics of Caenorhabditis elegans. Genetics, 1974, 77(1): 71~94.

[2] Sulston, J. E., Neuronal cell lineages in the nematode, Caenorhabditis elegans. Cold Spring Harb Symp Quant Biol, 1983, 48(2): 443~452.

[3] Sulston, J. E, Horvitz, H. R. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol, 1977, 56 (1): 110~156.

[4] Bao Z, Murray JI, Boyle T, Ooi SL, Sandel MJ, Waterston RH: Automated cell lineage tracing in Caenorhabditis elegans. Proc Natl Acad Sci USA 2006, 103(8):2707-2712.

[5] A semi-local neighborhood-based framework for probabilistic cell lineage tracing. Anthony Santella, Zhuo Du and Zhirong Bao. BMC Bioinformatics 2014, 15:217.

[6] Aydin Z, Murray JI, Waterston RH, and Noble WS: Using machine learning to speed up manual image annotation: application to a 3D imaging protocol for measuring single cell gene expression in the developing C. elegans embryo. BMC Bioinformatics 2010, 11:84.

[7] J.-H. Lee and C.-H. Won "Topology preserving relaxation labeling for nonrigid point matching", IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp.427 -432 2011.

[8] Q.X. Wu and D. Pairman, "A Relaxation Labeling Technique for Computing Sea Surface Velocities from Sea Surface Temperature," IEEE Trans. Geoscience and Remote Sending, vol. 33, no. 1, pp. 216-220, Jan. 1995.

[9] Chen, L., Chan, L., Zhao, Z., Yan, H.: A novel cell nuclei segmentation method for 3D C. elegans embryonic time-lapse images. BMC bioinformatics 14(1) (2013)32810.

[10] Dufour A, Shinin V, Tajbakhsh S, Guillen-Aghion N, Olivo-Marin JC, Zimmer C: Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. IEEE Trans Image Process 2005, 14(9):1396-1410.

[11] Dzyubachyk O, Jelier R, Lehner B, Niessen W, Meijering E: Model-based approach for tracking embryogenesis in Caenorhabditis elegans fluorescence microscopy data. Conf Proc IEEE Eng Med Biol Soc 2009, 1:5356-5359.

[12] Jaqaman, K. et al. Robust single-particle tracking in live-cell time-lapse sequences. Nat. Methods 5, 695–702 (2008).

[13] R. Sinkhorn, "A Relationship between Arbitrary Positive Matrices and Doubly Stochastic Matrices," The Annals of Math. Statistics, vol. 35, no. 2, pp. 876-879, 1964.