A NEW METHOD FOR ESTIMATING CELL LINEAGE: THEORY AND APPLICATION.

Atsushi Mochizuki, Makoto Koga and Yasumi Ohshima

Department of Biology, Kyushu University, Fukuoka 812-8581, Japan

E-mail: amochscb@mbox.nc.kyushu-u.ac.jp

1. INTRODUCTION

A multicellular organism is composed of cells that are originated in a single cell, a fertilized egg. Development proceeds as cell divisions and occasional cell differentiations. In the early development of some invertebrates, the rule of cell division is invariant between embryos. The genealogical relationship of cells is summarized as a "cell lineage," which describes the developmental pattern of an embryo from the view of cell divisions.

In this paper, I show the systematic estimate of cell lineage from the distributions of marked cells at an observation stage, assuming that markers are introduced randomly in earlier developmental stages. The idea was developed by one of the authors and shown in the previous paper [1]. If only a small number of marker insertions occur in an experiment, the marked cells at the observation stage are inferred to be the descendants of a few marker-inserted cells at an earlier stage. If in addition the number of replicate experiments is sufficiently large, we can estimate the cell lineage leading to cells at the observation stage.

When many cells are included at the observation stage, the total number of possible cell lineages becomes enormous. Then, the faster method for estimation, clustering method, was developed in which topology (i.e. tree shape) of cell lineage is reconstructed by sequential merging of pairs of cells that have the highest correlation in observed marker labelling. This method enables us to search for the correct topology of cell lineage very quickly even if many cells are included. The ability of estimate the correct lineage topology is examined for many examples of hypothetical lineages by computer simulations. Application to two organisms are also shown.

2. SYMBOLIC REPRESENTATION

Let us consider (a part of) the body of an organism which is made of distinguishable N cells at a stage in which observation occurs. These N cells are called observed cells. Suppose that, in the development of this organism, the rule of cell division is the same between embryos. I assume that these N observed cells are produced from a single cell by N-1 cell divisions. I ignore the possibility of cell-fusion.

Suppose an experimental procedure, by which the marker would be introduced into a small number of cells in an embryo at any stage earlier than the observation stage. One experiment results in a distribution of marked cells among N cells at the observation stage. The number of markers introduced into each embryo is variable between experiments. This experiment is iterated for K times using different embryos, expecting that sufficiently different ways of marker insertions are included if K is sufficiently large. I assume that the cell lineage is not affected by the presence of the marker. Further, I assume that the marker is inherited to both daughter cells without failure. Genetic markers may not be diluted. Chemical markers should be diluted but may be still detectable even if diluted by cell divisions until the observation stage. In total, K distributions of marked cells are used for estimating cell lineage.

I indicate marked and unmarked cells by 1 and 0, respectively. Let x_i^m be the state of the ith observed cell in the mth experiment ($1 \le m \le K$ and $1 \le i \le N$). The result of the whole set of experiments is summerized as a (K, N)-matrix $\mathbf{A}$. The ith column of the matrix $\mathbf{A}$ is the state of the ith observed cell in the whole experiments:

$$\mathbf{S}_i = \left(x_i^1, x_i^2, \ldots, x_i^K\right)^T, \tag{1}$$

where T indicates transposition of the vector.

Let $\{a, b\}$be the symbol for the parent cell which is to produce cell a and cell b after division. We can infer the ancestral cell state (presence or absence of marker) as that the state of parent cell $\{a, b\}$ is 1 if and only if both daughters are 1. This means that $x_{\{a,b\}}^m = 1$ if and only if $x_a^m = 1$ and $x_b^m = 1$, and hence $x_{\{a,b\}}^m = x_a^m x_b^m$. If we consider all the K replicate experiments, we have:

$$\mathbf{S}_{\{a,b\}} = \left(x_a^1 x_b^1, x_a^2 x_b^2, \ldots, x_a^K x_b^K\right)^T. \tag{2}$$

Suppose that the mth component of $\mathbf{S}_{\{a,b\}}$ is 0 but the mth component of $\mathbf{S}_a$ is 1. This means that a marker was inserted into cell a after it is divided from $\{a, b\}$. Hence the difference in the number of 1's among the components between $\mathbf{S}_a$ and $\mathbf{S}_{\{a,b\}}$ indicates the number of marker-insertions into cell a after the division of cell $\{a, b\}$ in the N experiments. Total number of marker-introductions in cell a is $\left\|\mathbf{S}_a - \mathbf{S}_{\{a,b\}}\right\|_1$, where $\|\bullet\|_1$ indicates absolute-sum norm ($\|(x,y,z)\|_1 = |x| + |y| + |z|$). Similarly, the number of marker-insertions into cell b after the division of cell $\{a, b\}$ can be calculated.

The above inference holds for each cell-division in the cell lineage. If we fix a topology of cell lineage, we can determine all the states of cells in the intermediate stages of the cell lineage following the procedure described by Eq. (2). By iterating this procedure for all the cell-divisions

in cell lineage sequentially from the last-divided cells to earlier-divided cells, we can determine the states of all the cells in the lineage. The required number of marker-insertion to explain the marker distribution at observation stage is also calculated.

The number of possible topologies of cell lineage with N observed cells is known to be $3 \cdot 5 \cdot \cdots (2N-3)$. When N is large, this number becomes very large and makes impossible to examine all the possible topologies. For example, it exceeds thirty million if N is equal to 10. Then, I introduce an alternative and a more practical method to find the cell lineage with the minimum number of marker insertions. In each step, we search for a pair of cells produced from a single parent cell by division by choosing the pair with the highest correlation of marker distributions. This method is based on the intuition that the two cells marked together in many experiments are likely to be the pair produced from a late cell division. The detail of the algorithm is shown in Mochizuki [1].

3. THE SUCCESS RATE

I examined the efficiency of the clustering method in estimating the correct cell lineage by computer simulation. I first generated a hypothetical topology of cell lineage, and artificial experiments of marker-insertion were carried out based on this cell lineage. Then the cell lineage was estimated by using the clustering method from the marker-distributions at the observation stage, and examined whether the estimated topology was the correct one. More than one markers might be inserted independently into one embryo in each experiment. The probability that n markers were inserted in an experiment (i.e. in the whole cell lineage) was assumed to follow a Poisson distribution with mean λ ($P(n) = \lambda^n e^{-\lambda}/n!$).

I calculated the "success rate" of estimation. It is 0 if the correct topology was not included in the set of estimated topologies. If the correct topology was included, then the success rate is the inverse of the number of candidates. This values are averaged over different topologies and over different sets of data.

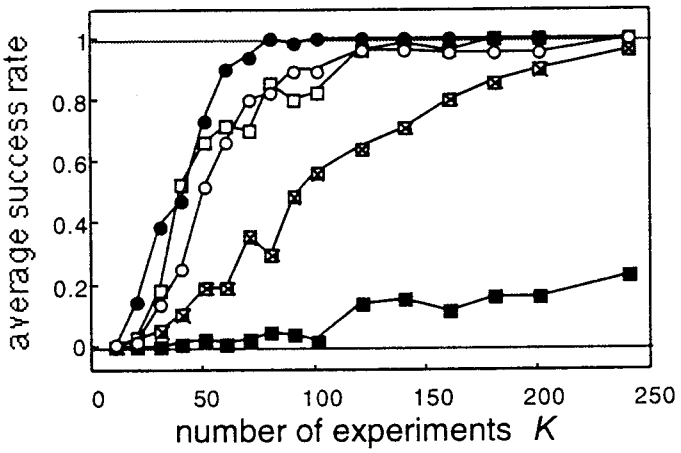

Figure 1 The averaged success rate of estimation.

Fig. 1 shows the value when N is 20. The horizontal axis is the number of experiments K. The value averaged for different topologies and for different sets of data is shown in vertical axis. Each line indicates the result of the cases in which mean number of insertion λ is equal to 1 (-○-), 3 (-●-), 5 (-□-), 9 (-⊠-), and 13 (-■-), respectively. The curve of success rate increased with K and the value finally reached 1 unless λ is too large (λ =13) or too small (λ =1). This implies that the correct cell lineage could be always estimated as the single candidate topology for a sufficiently large K. There was optimal value of λ for estimating correct topology, which is about 5 when N is 20. The optimal λ seems to depend on N.

4. Reconstruct Cell Lineage of Ascidian from Experimental Data

In this section, I try to reconstruct the cell lineage of ascidian based on the marker-distributions at a single observation stage, which was shown in the study by Nishida & Sato [2, 3] and Nishda [4]. They studied cell lineage of whole organism of ascidian in early development by using intercellular marker HRP. In their study, each cell was identified by direct observation, and one of the identified cells at intermediate stage was marked by injection of HRP [2, 3, 4].

From these data, I constructed the data of marker-distributions at tailbud-stage showing experimental results. In the estimation, the information about the identity of marker-inserted cell for each distribution was neglected. First, I determined the "unit regions" in the body at an observation stage, which corresponds to observed cells in the above mathematical modelling. Then for each photograph, I indicated whether each unit was stained or not by 1 or 0, respectively. These are regarded as the data $\mathbf{A} = [x_i^m]$. Then by applying the clustering method to $\mathbf{A}$, the topology of the cell lineage was estimated.

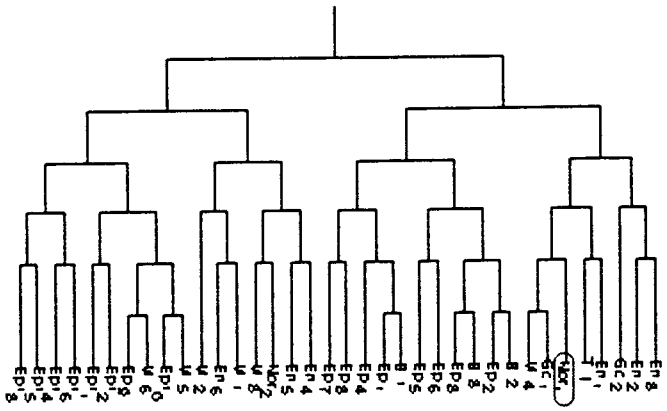

Figure 2 The estimated lineage of ascidian.

The obtained topologies are shown in Fig. 2. They are the same as the cell lineage derived previously by Nishida & Sato [2, 3] and Nishda [4].

5. Reconstruct Cell Lineage of *C. elegans* from Experimental Data

We tried to reconstruct a part of cell lineage of actual organism, nematode, *C. elegans* by using the data obtained from random marker-introduction. By comparing the lineage of *C. elegans* previously determined by direct observation [5, 6], we can confirm reliability of the clustering method. We focus on the cell lineage of intestine made from 20 cells, where each cell is large and easy to observe.

The mechanism of random marker-introduction is realized by the extrachromosomal array in the nematode cell. We can make lines of nematode experimentally which carry some copies of the circular gene beside the chromosome. The array is slightly unstable and may be lost occasionally from each cell during somatic cell division. We prepared a gene construct containing GFP (green fluorescent protein) gene with a promoter sequence that is activated in intestine cells. Then a line carrying the gene construct in extrachromosomal array was prepared. The intestine cells without fluorescence are the descendant of the cell(s) which lost the array. Then just by observing the distribution of the lack of the fluorescence by using microscope, we obtain the data of distribution of marked cells at the observation stage. The observation stage was set to stage L1 or L2, where the cell number is the same as that of the adult nematode but identifying each cell is much easier than at adult stage.

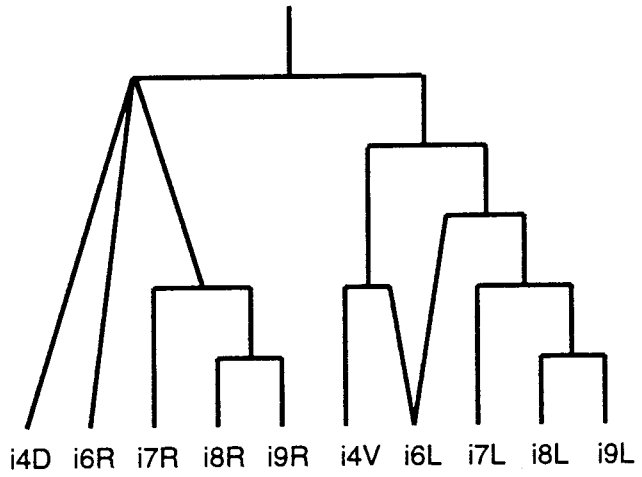

Figure 3 The estimated partial lineage of *C. elegans*.

In this analysis, 40 embryos having some lack of fluorescence were observed. Fig. 3 shows a part of cell lineage of intestine estimated by clustering method using a part of the observation data. The full result will be shown in the presentation at the conference.

REFERENCES

[1] MOCHIZUKI, A. Estimating cell lineage from distributions of randomly introduced markers. *J. theor. Biol* **197**: (1999), 227-245.

[2] NISHIDA, H. & SATO, N. Cell lineage analysis in ascidian embryo by intracellular injection of marker enzyme. I. Up to the eight-cell stage. *Develop. Biol.* **99**: (1983), 382-394.

[3] NISHIDA, H. & SATO, N. Cell lineage analysis in ascidian embryo by intracellular injection of marker enzyme. II. The 16- and 32- stage. *Develop. Biol.* **110**: (1985), 440-454.

[4] NISHIDA, H. Cell lineage analysis in ascidian embryo by intracellular injection of marker enzyme. III. Up to the tissue restricted stage. *Develop. Biol.* **121**: (1987), 526-541.

[5] SULSTON, J. E. & HORWITZ, H. R. Post-embryonic cell lineage of the Nematode, *Caenorhabditis elegans. Develop. Biol.* **56**: (1977), 110-156.

[6] SULSTON, J. E., SCHIERENBERG, E., WHITE, J. G. & THOMSON, J. N. The embryonic cell lineage of the Nematode, *Caenorhabditis elegans. Develop. Biol.* **100**: (1983), 64-119.