Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, Miami, Florida, USA, January 18-20, 2006 (pp227-232)

The Statistical Analysis of Longitudinal Clonal Data on Oligodendrocyte Generation

Ollivier Hyrien[a,*], Margot Mayer-Pröschel[b], Mark Noble[b], and Andrei Yakovlev[a]
[a]Department of Biostatistics and Computational Biology,
[b]Department of Biomedical Genetics,
University of Rochester Medical Center, Rochester, NY 14642, USA

Abstract: This paper is concerned with the analysis of the generation of oligodendrocytes from oligodendrocyte-type 2 astrocytes progenitor cells in tissue culture using longitudinal observations on clonal growth. Our approach is based on a multi-type age-dependent branching process, and the statisical analyses are conducted using a simulated pseudo maximum likelihood estimator. We evaluate the performance of our method in a simulation study, and apply it to experimental data on the growth of clones of oligodendrocytes.

Key-Words: Branching process; Pseudo likelihood inference; Multi-type cell populations;

1. Introduction

The quantitative analysis of the processes of division and differentiation of cultured progenitor cells, where cell proliferation is observed at the clonal level, has been extensively investigated in past studies [1-10]. Despite the fact that the events of interest are not directly observable under these experimental conditions which provide only counts of progenitor cells and differentiated cells, these references have shown that important characterisitcs of cell kinetics can be accurately estimated from such observed data. The various methods proposed in these publications were based on the theory of multi-type age-dependent branching processes, and the authors used computer-intensive techniques to conduct the desired statistical analyses. The proposed methodologies were succesfully applied to analyze the generation of oligodendrocytes from their oligodendrocytes type-2 astrocytes (subsequently abbreviated as O-2A) progenitor cells cultured *in vitro*, and several characteristics of the processes of division, differentiation and death of this cell system have been quantified.

These analyses were all performed on clonal data yielded by experiments where the composition of every clone was examined only once before being discarded. This type of data will be subsequently referred to as independent clonal data in contrast to longitudinal clonal data. The present paper focuses on a different situation where each clone is examined at multiple time points, thus providing longitudinal measurements on clonal growth. These data provide more information on the processes underlying the generation of cellular clones, allowing potentially for the construction of more complex and more realistic mathematical models of cell proliferation. The goal of the present paper is to apply our methodology to analyze longitudinal clonal data on oligodendrocyte generation from O-2A progenitor cells cultured *in vitro*. We consider modeling the growth of clones composed of these cells using the multi-type age-dependent branching process proposed in [8]. This process is an extension of a multi-type Bellman-Harris branching process which allows for dissimilar distributions for the time to division and the time to differentiation of O-2A progenitor cells. In order to fit this model to longitudinal clonal data, we suggest using the simulated pseudo maximum likelihood approach proposed in [8,9,10] for age-dependent branching processes observed at discrete time points. This article is illustrated using an experimental data set on the generation of oligodendrocytes from cultured O-2A progenitor cells. The properties of our method for longitudinal clonal data are next investigated in a simulation study.

2. A Model of O-2A Proliferation

We consider the multi-type branching process proposed in [8]) to represent the proliferation of O-2A progenitor cells and their ultimate transformation into oligodendrocytes. The model is based on the following assumptions:

(1) The process begins with a single progenitor cell cultured at time $t = 0$. At the end of the kth mitotic cycle, every progenitor cell either gives rise to two progenitor cells with probability p_k, or it differentiates into one oligodendrocyte with probability $1 - p_k$. The division probability p_k is a decreasing function of the mitotic cycle number $k \in \{1, 2, ...\}$, where $k = 1$ corresponds to the first cycle (division) after plating. We assume the following form of dependence of p_k on k:

$$p_k = a + bc^k, \qquad k \geq 1.$$

(2) Since progenitor cells appear to have a very high survival rate, the model assumes that O-2A progenitor cells do not die during the time of the experiment. The death of oligodendrocytes normally begins on day 7 which is beyond the range of our data. Therefore, we assume that oligodendrocytes neither divide nor die.

(3) The time to division of any O-2A progenitor cell follows a gamma distribution with mean m_1 and variance σ_1^2.

(4) The time to differentiation of any O-2A progenitor cell follows a gamma distribution with mean m_2 and variance σ_2^2.

(5) Progenitor cells do not migrate out of the field of observation.

(6) The usual independence assumptions regarding the evolution of age-dependent branching processes are adopted.

The set of (unknown) parameters of the model is the vector $\theta = (a, b, c, m_1, \sigma_1^2, m_2, \sigma_2^2)'$. We define the 2-dimensional stochastic process $\mathbf{Z}(t, \theta) = \{Z_1(t, \theta), Z_2(t, \theta)\}'$ where $Z_1(t, \theta)$ and $Z_2(t, \theta)$ represent the number of O-2A progenitor cells and the number of oligodendrocytes at time t in a clone generated by the branching process defined above.

3. Statistical Inference

3.1 Longitudinal Clonal Data Longitudinal clonal experiments consist in plating O-2A progenitor cells in a growth medium at a sufficiently low density so they generate separate clusters of cells. such clusters will be referred to as cell clones in what follows. The experimentalist examines the composition of each resulting clone at several points in time to obtain the number of O-2A progenitor cells and the number of oligodendrocytes. Thus, the observations can be represented as a set of vectors $\mathbf{Y}_i = (Y_{i1}, ..., Y_{im_i})'$, $1 \leq i \leq n$, where $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2})$ and where Y_{ij1} and Y_{ij2} denote the number of O-2A progenitor cells and the number of oligodendrocytes counted at time t_{ij} in the ith clone. We write $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$ for the times of observation. It is assumed that the observations $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are independent random vectors (r.v.) generated by the branching process formulated in Section 2. The true value of the vector θ is denoted by θ_0.

3.2 Moment-based Estimation

Because of the complexity of our model, neither the distribution nor the moments of the process $\mathbf{Z}(t, \theta)$ have known explicite analytical expressions. As a result, the statistical analysis of clonal data can only be achieved through approximations of these quantities. This paper considers simulation-based approximations.

Simulated maximum likelihood estimation was investigated in [5] in the context of independent clonal data. In brief, the proposed method consisted in simulating a large number of independent clones from a computer version of the model. The distribution function of the process $\mathbf{Z}(t, \theta)$ was then approximated using its empirical counterpart. However, when the composition of some observed clones match none of the composition of their simulated counterparts, the resulting simulated likelihood function is identical to zero, and the simulated log-likelihood $L_S(\theta)$ is undefined. These mismatches, which were encountered in the independent setting [5], make it difficult to compute the parameter estimate. Since they are even more likely to occur as m_i increases, they may become a serious burden when analyzing longitudinal clonal data using the simulated maximum likelihood approach.

Alternative methods of estimation that will resolve the mismatching problem in the context of longitudinal clonal data can be designed from simulation-based approximations of moments of the observed quantities. The simulated pseudo maximum likelihood estimator proposed in [8,9,10] for discretely observed branching processes provides such an example. This method is solely based on the mean vector and the variance-covariance matrix of the observed cell counts. Thus, for each i, let us introduce the $Kn_i \times 1$ vector

$$\mu_i(\theta) = \{\mathbf{m}(t_{i1}, \theta)', ..., \mathbf{m}(t_{in_i}, \theta)'\}', \quad 1 \leq i \leq n$$

where $\mathbf{m}(t_{ij}, \theta) = E\{\mathbf{Z}(t_{ij}, \theta)\}$ is a vector representing the mean numbers of cells of O-2A progenitor cells and the mean number of oligodendrocytes counted at time t_{ij}, given the parameter value θ. Let $\Omega_i(\theta)$ denote the associated $Kn_i \times Kn_i$ variance-covariance matrix:

$$\Omega_i(\theta) = \begin{pmatrix} V(t_{i1}, t_{i1}, \theta) & ... & V(t_{i1}, t_{in_i}, \theta) \\ \vdots & \ddots & \vdots \\ V(t_{in_i}, t_{i1}, \theta) & ... & V(t_{in_i}, t_{in_i}, \theta) \end{pmatrix},$$

where $\mathbf{V}(t_{ij_1}, t_{ij_2}, \theta) = \text{cov}\{Z\mathbf{Z}(t_{ij_1}, \theta), \mathbf{Z}(t_{ij_2}, \theta)\}$. Because $\mu_i(\theta)$ and $\Omega_i(\theta)$ have no explicite analytical expressions, we will resort to simulations to approximate their values as described below.

Let $\mathbf{Z}^{\star,s}(t, \theta) = \left\{Z_1^{\star,s}(t, \theta), Z_2^{\star,s}(t, \theta)\right\}$, $1 \leq s \leq S$, be S independent random vectors, each of them representing the number of progenitor cells and the number of oligodendrocytes counted at time t in the sth clone simulated using a computer version of the assumed model, and given the parameter vector θ. Let $\mathbf{Y}_i^{\star,s}(\theta) = \{\mathbf{Z}^{\star,s}(t_{i1}, \theta), ..., \mathbf{Z}^{\star,s}(t_{im_i}, \theta)\}$. The random vectors $\mathbf{Y}_i^{\star,s}(\theta_0)$, $1 \leq s \leq S$, can be considered as i.i.d. copies of the observed vectors Y_i. The mean vectors $\mu_i(\theta)$ and the variance-covariance matrices $\Omega_i(\theta)$ can be approximated by their empirical estimators $\mu_i^S(\theta)$ and $\Omega_i^S(\theta)$ defined as

$$\mu_i^S(\theta) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{Y}_{i,s}(\theta) \tag{1}$$

and

$$\Omega_i^S(\theta) =$$

$$\frac{1}{S-1} \sum_{s=1}^{S} \{\mathbf{Y}_{i,s}(\theta) - \mu_i^S(\theta)\}\{\mathbf{Y}_{i,s}(\theta) - \mu_i^S(\theta)\}'. \tag{2}$$

By the Strong Law of Large Numbers, $\mu_i^S(\theta)$ and $\Omega_i^S(\theta)$ converge almost surely to $\mu_i(\theta)$ and $\Omega_i(\theta)$ as $S \to \infty$. The simulated pseudo likelihood $G_n^S(\theta)$ is defined as

$$G_n^S(\theta) = -\sum_{i=1}^{n} C_i^S(\theta) \tag{3}$$

where $C_i^S(\theta)$ represents the contribution of the ith clone given by

$$C_i^S(\theta) = \{\mathbf{Y}_i - \mu_i^S(\theta)\}'\Omega_i^S(\theta)^{-1}\{\mathbf{Y}_i - \mu_i^S(\theta)\}$$

$$+ \log|\Omega_i^S(\theta)|.$$

The simulated pseudo maximum likelihood estimator is a vector $\hat{\theta}_n^S$ which maximizes $G_n^S(\theta)$: $\hat{\theta}_n^S = \text{ArgMax}_{\theta\in\Theta} G_n^S(\theta)$. Hyrien [10] established asymptotic properties of the simulated pseudo maximum likelihood estimator for the traditional multi-type Bellman-Harris branching process, and they remain true in the present setting as well. In brief, the estimator converges almost surely to the pseudo maximum likelihood estimator, $\hat{\theta}_n$, as S tends to infinity, which maximizes a pseudo likelihood $G_n(\theta)$ constructed from the exact mean vectors and variance-covariance matrices of the observed cell counts. Under certain regularity conditions, the estimator $\hat{\theta}_n$ converges almost surely to θ_0 as n tends to infinity, and in large samples, it is approximately normally distributed with zero mean and variance-covariance matrix $\Sigma(\theta_0) = I(\theta_0)^{-1}J(\theta_0)I(\theta_0)^{-1}/n$, where $I(\theta_0) = \lim_{n\to\infty} \text{E}_{\theta_0}\{\nabla^2_{\theta\theta'} G_n(\theta_0)\}/n$ and $J(\theta_0) = \lim_{n\to\infty} \text{E}_{\theta_0}\{\nabla_\theta G_n(\theta_0)\nabla_\theta G_n(\theta_0)'\}/n$. These properties are also approximately satisfied by $\hat{\theta}_n^S$ if S is large enough.

The estimating function $G_n^S(\theta)$ fluctuates randomly because the moments $\mu_i(\theta)$ and $\Omega_i(\theta)$, $1 \leq i \leq n$, are approximated by Monte Carlo integration. In order to compute the simulated pseudo maximum likelihood estimator, one can implement the sample path approach as advocated in [8,9,10]. Further, since branching stochastic processes are jump processes, the simulated pseudo likelihood is not continuous over the parameter space Θ, but rather a step function. It is therefore better to use optimization

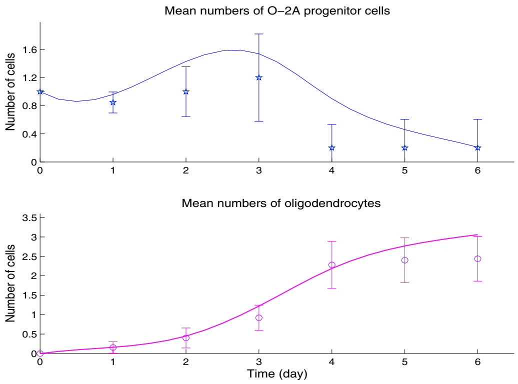

Figure 1: Empirical and fitted mean number of O-2A progenitor cells and oligodendrocytes over time.

algorithms that do not require differentiating the estimating function.

The variance-covariance matrix can be estimated using a parametric bootstrap, and test of nested hypotheses can be carried out using the Monte Carlo Wald test proposed in [8].

4. An Application

In this application, the model is used to obtain quantitative insights into the process of clonal growth and differentiation of oligodendrocyte-type 2 astrocyte (O-2A) progenitor cells cultured *in vitro*. Experimental data were obtained from experiments carried out on O-2A progenitor cells isolated from the rat optic nerve. Twenty five O-2A progenitor cells were plated in separate flasks in a culture medium added with thyroid hormone. The composition of each of the 25 clones was recorded at the following time points: 24, 48, 72, 96, 120 and 144 hours after the start of the experiment. These observations represent longitudinally observed pairs of counts of O-2A progenitor cells and oligodendrocytes.

This model of clonal development has proven to be in good agreement with the observed dynamics of O-2A progenitor cells and oligodendrocytes in several sets of experimental data. However, the longitudinal observation process provides much more information on the system under study than the earlier used data on independent cell counts (with each clone being scored only once) at different time points.

We fit our model using the simulated pseudo likelihood approach. For that purpose, we computed the empirical estimators of $\mathbf{m}(t,\theta)$ and $V(u,t,\theta)$ with the number of simulated clones S gradually increasing during the process of optimization. The ultimate parameter estimates were obtained with $S = 100,000$ simulated clones. To maximize $G_n^S(\theta)$, we employed a random search algorithm combined with the sample path methodology. The algorithm was launched several times to assess the obtained values as well as to guard against the random noise arising from the simulations. The estimate of the parameter vector $\theta = (a, b, c, m_1, \sigma_1, m_2, \sigma_2)'$ is given in Table 1.

The parameter estimates are in quite good agreement with those previously obtained in our past pubications. It is also interesting to note that the mean time to differentiation still appears to be quite close to the mean time to division, but there is a marked difference between the corresponding variances. The standard error of the simulated pseudo maximum likelihood estimates reported in Table 1 was obtained by the parametric bootstrap with 400 bootstrap samples.

Parameter	m_1	σ_1	m_2	σ_2	a	b	c
Estimate	27.6	10.2	24.6	25.2	0.13	11.4	0.05
Std error	2.20	2.24	2.84	4.47	0.02	1.95	0.01

Table 1: Parameter estimates

	Parameter	m_1	σ_1	m_2	σ_2	a	b	c
	True values	25.0	10.0	30.0	28.3	0.20	1.0	0.5
n=25 clones	Average	24.5	9.4	31.4	28.2	0.18	1.10	0.5
	Std. dev.	1.2	2.3	3.3	4.7	0.04	0.22	0.05
n=50 clones	Average	25.1	9.9	30.1	27.4	0.19	1.05	0.49
	Std. Dev.	1.3	2.6	2.7	4.5	0.02	0.15	0.05
n=150 clones	Average	24.9	9.9	30.1	28.1	0.20	1.00	0.50
	Std. Dev.	0.6	1.1	1.8	3.1	0.02	0.08	0.03

Table 2: Results of the simulation study

Figure 1 presents the mean numbers of O-2A progenitor cells and oligodendrocytes and the corresponding model-based estimates. As is seen from the figure, the model captures the pattern of the mean number of cells, but the fit is not very good around day 4 for the mean number of O-2A progenitor cells. This lack of fit indicates that our model can be further improved based on longitudinal clonal data to gain a better understanding of the generation of oligodendrocytes from cultured O-2A progenitor cells.

5. A simulation Study

We conducted a simulation study to investigate finite sample properties of the simulated pseudo maximum likelihood estimator in the context of the proposed model. The estimates of model parameters reported in Table 1 were used to generate several data sets in accordance with the following experimental design: for each data set, a total number of n independent clones were simulated and the numbers of progenitor cells and oligodendrocytes in each of these n clones were counted each day from day 1 to day 6, in exactly the same way as they were counted in the biological experiment. To assess the quality of the estimation procedure, we used 100 replicates for the following sample sizes: $n = 25$, 50 and 100. The moments of the observed numbers of cells were estimated from $S = 25,000$ simulated clones. The results reported in Table 2 show that the simulated pseudo maximum likelihood estimation procedure combined with the sample path method and random search performs fairly well even with small samples.

6. Discussion

The present paper considered an application of a previously proposed branching stochastic process to the analysis of longitudinal clonal data on the generation of oligodendrocytes from cultured O-2A progenitor cells. Because of potential mismatches encountered with the simulated maximum likelihood approach, the simulated pseudo maximum likelihood estimator remains a method of choice in the considered setting. Our simulation study indicated that this estimator can be expected to perform well in finite samples of longitudinal clonal data, even with a relatively low number of replicates (say 25 clones).

Furthermore, we applied our method to a new data set yielded by a longitudinal clonal experiment where each cellular clone was observed at 6 different points in time. The branching stochastic process considered here was found to provide a good fit to independent clonal data in Hyrien et al (2005a), and this model suggested that the time to division and the time to differentiation of O-2A progenitor cells followed dissimilar distribution functions. The new analysis of longitudinal data reported in the present paper supported the latter conclusion, but it also indicated that this model may not be as appropriate to describe longitudinal clonal data (since the resulting fit was not as good) as it was when fitted to independent clonal data. In order to improve our statistical analyses, the model could be further developed. Many such improvements could be considered, including for example the so-called clonal inheritance assumption. At the moment, it is however unclear by looking at clonal data which part of the model needs to be modified, and time-lapse experiments currently underway will be helpful in this regards. We believe that longitudinal clonal data are informative enough to sustain further modeling efforts.

Acknowledgement

This research was supported by NIH/NINDS grants NS39511 and NS37166.

References

[9] Yakovlev, A.Y., Mayer-Pröschel, M., and Noble, M. (1997). A stochastic model of oligodendrocyte generation in culture. *Cell Proliferation* **30**, 244.

[10] Yakovlev, A.Y., M. Mayer-Pröschel and M. Noble (1998) A stochastic model of brain cell differentiation in tissue culture, *Journal of Mathematical Biology*, **37** 49-60.

[11] Yakovlev, A.Y., M. Mayer-Pröschel and M. Noble (2000) Stochastic formulations of a clock model for temporally regulated generation of oligodendrocytes *in vitro*, *Mathematical and Computer Modelling*, **32** 125-137.

[4] von Collani, E., Tsodikov, A., Yakovlev, A., Mayer-Pröschel, M., and Noble, M., (1999). A random walk model of oligodendrocyte generation in vitro and associated estimation problems. *Mathematical Biosciences* **159**, 189.

[5] Zorin, A.A., Yakovlev, A.Y., Mayer-Pröschel, M., and Noble, M. (2000). Estimation problems associated with stochastic modeling of proliferation and differentiation of O-2A progenitor cells in vitro. *Mathematical Biosciences* **167**, 109-121.

[6] Boucher, K., Yakovlev, A.Y., Mayer-Pröschel, M., and Noble, M. (1999). A stochastic model of temporarily regulated generation of oligodendrocytes in vitro. *Mathematical Biosciences* **159**, 47-78.

[7] Boucher, K., Zorin, A., Yakovlev, A.Y., Mayer-Pröschel, M. and Noble, M. (2001). An alternative stochastic model of generation of oligodendrocytes in cell culture. *Journal of Mathematical Biology* **43**, 22-36.

[8] Hyrien, O., Mayer-Pröschel, M., Noble, M., Yakovlev, A., 2005a. A stochastic model to analyze clonal data on multi-type cell populations. *Biometrics*, **61**, 199-207.

[9] Hyrien, O., Mayer-Pröschel, M., Noble, M., Yakovlev, A., 2005b. Estimating the lifespan of oligodendrocytes from clonal data on their development in cell culture. *Mathematical Biosciences*, **193**, 255-274.

[10] Hyrien, O. (2005). A pseudo maximum likelihood estimator for discretely observed multi-type Bellman-Harris branching processes. In revision.