Table 2. Estimated Parameter Values in the Gamma-Frailty Model

A. Homogeneous multi-target model

Number of targets	$\hat{\beta}$	$\hat{\rho}$	$\widehat{RR}_e$	$\widehat{RR}_m$	$\widehat{RR}_l$	*AIC*
11	0.3088 (0.2937, 0.3240)	1.0	0.914	0.912	0.859	996.11

B. Heterogeneous multi-target model with single stem cell assumption

Number of targets	$\hat{\beta}$	$\hat{\rho}$	$\widehat{RR}_e$	$\widehat{RR}_m$	$\widehat{RR}_l$	*AIC*
10	0.2362 (0.2343, 0.2380)	1.067	0.910	0.912	0.856	996.29
20	0.2313 (0.2307, 0.2319)	1.075	0.912	0.912	0.859	993.06
30	0.2315 (0.2311, 0.2319)	1.075	0.913	0.912	0.859	992.98
40	0.215 (0.2312, 0.2318)	1.075	0.913	0.912	0.859	992.98

C. Heterogeneous multi-target model with multiple stem cell assumption

Number of stem cells	Number of targets	$\hat{\beta}$	$\hat{\rho}$	$\widehat{RR}_e$	$\widehat{RR}_m$	$\widehat{RR}_l$	*AIC*
2	10	0.2391 (0.2386, 0.2397)	1.157	0.912	0.912	0.859	993.08
	15	0.2393 (0.2389, 0.2397)	1.156	0.912	0.912	0.859	993.03
	20	0.2393 (0.2390, 0.2396)	1.156	0.912	0.912	0.859	993.03
3	8	0.2468 (0.2463, 0.2472)	1.248	0.912	0.912	0.859	993.04
	10	0.2468 (0.2465, 0.2471)	1.247	0.912	0.912	0.859	993.03
	12	0.2468 (0.2465, 0.2471)	1.247	0.912	0.912	0.859	993.03
4	10	0.2544 (0.2541, 0.2546)	1.349	0.912	0.912	0.859	993.08
	12	0.2543 (0.2540, 0.2546)	1.349	0.912	0.912	0.859	993.08

Note: $\hat{\sigma}^2$=0.006
Values in parentheses are the 95% confidence intervals

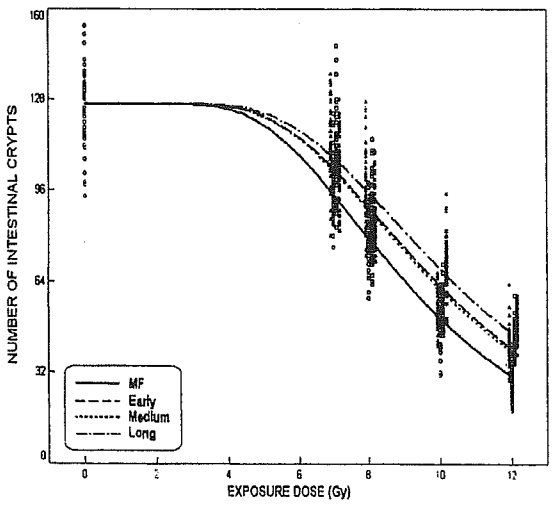

Fig. 2. Curve of the density of the intestinal crypt after an exposure event based on the Poisson regression model according to the fermented-stage of miso, with mice fed with a commercial diet of MF used as controls. The survival rate of crypts of mice fed long-term fermented miso has a higher rate indicated by the slope of the curve slightly decreasing as compared with the others. On the other hand, the short-term and medium-term fermentations confer almost the same level of protection on the crypts after exposure. In the scatter plot results for the mice exposed to 7, 8, 10 or 12 Gy of X-rays after being fed a commercial diet of MF marked by a circle or a diet supplemented with miso fermented for a short-, medium-, or long-term marked by a triangle, square, and asterisk respectively.

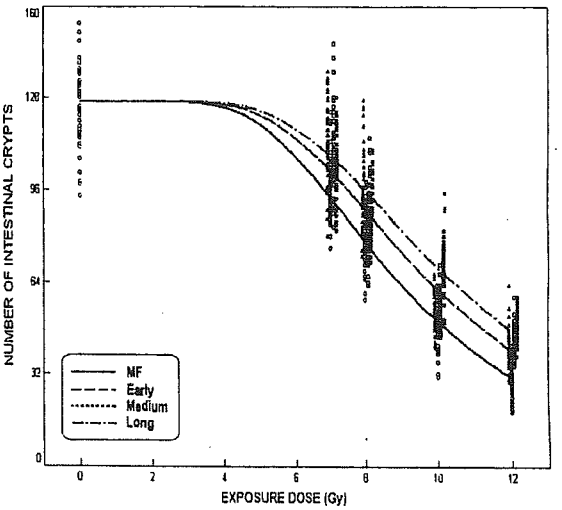

Fig. 3. Curve of the density of the intestinal crypt after an exposure event based on the gamma-frailty model according to the fermented-stage of miso, with mice fed a commercial MF diet used as controls. The survival curve of the crypts of mice fed long-term fermented miso has a slightly decreasing slope, indicating that the crypt-survival rate in this group was higher than in the other groups. On the other hand, the short-term and the medium-term fermentations confer exactly the same level of protection on the crypts after exposure. In the scatter plot results for the mice exposed to 7, 8, 10 or 12 Gy of X-rays after being fed a commercial diet of MF marked by a circle or a diet supplemented with miso fermented for a short-, medium-, or long-term marked by a triangle, square, and asterisk respectively.

ted model measurement were significantly lower when the gamma-frailty model was applied than when the Poisson regression model was used. For protecting the crypts after exposure, both the Poisson regression model and gamma-frailty model yielded similar results on short-term and medium-term fermented miso, as shown by the similar values of the relative risk corresponding to the fermentation terms $\widehat{RR}_e$ and $\widehat{RR}_m$, respectively. On the other hand, the relative risk values of the long-term group ($\widehat{RR}_l$) were a little lower than the others, indicating significant protection of the crypts against the exposure effects (see Table 1 and Table 2). Furthermore, from a graphical point of view, the survival curve of the long-term group has a slightly decreasing slope, which means that the rate of crypt survival of this group is higher than that of the other groups (see Fig. 2 and Fig. 3). Moreover, these results show that the gamma-frailty model based on assumed heterogeneity in the target size, as indicated by the values of the heterogeneity index, is more suitable for application to such empirical data, in which the number of targets was 30 genes and the AIC value was 992.98. Regarding the number of stem cells in the crypt, it was suggested that the fitted model could be obtained when m=3 and there were at least 10 genes, as indicated by the AIC value of 993.03.

DISCUSSION

Results of Data Analysis

As mentioned in the previous section, the results showed that the dose-incidence curves reached a plateau at about 3 dead cells per crypt section in the mouse small intestine. This result is close to the result reported by Hendry et al[6] of about 3 to 4 dead cells per crypt section. Furthermore, they pointed out that the production of apoptotic cells by low doses of gamma-rays was independent of the dose rate between 0.27 and 450 cGy per min. Moreover, Hendry and Potten[7] reported that the cells that die via apoptosis represent a very sensitive subpopulation of about 6 cells per crypt that may or may not be clonogenic. Fujikawa et al[5] similarly reported that 5.1 ± 0.3 somatic crossing-over mutations were induced by X-rays in *Drosophila melanogaster* and Takai et al[20] estimated that 4.3 ± 0.6 such mutations were induced by X-rays in medaka fish (*Oryzias latipes*).

Identifiability

Application of a gamma-frailty multi-target model to this experimental exposure data revealed that the survival curves flattened out when the number of targets was more than 10, as indicated by the relatively stable AIC values. This may indicate either that the model is less sensitive in identifying cell changes in more than 10 targets, or that the cell changes have no significant effect on the model. Furthermore, the index value ρ related to the survival rate indicated that as the exposure dose (D) approached infinity, the number of targets k does not affect the change of survival rate when the index value ρ is greater than 1. On the contrary, when ρ equals 1, the survival rate of k targets tends to be k times the survival rate of one target (*see Proposition 1 in Appendix A*).

Related Topics

There is a long history of attempts to establish a theoretical model of exposure-induced cell changes. The multi-stage model proposed by Armitage and Doll[1] based on the hypothesis of Fisher and Hollomon[4] has been used in biomedical fields for more than fifty years. This hypothesis assumed that carcinogenic transformation of cells in a tissue requires that independent changes occur in six or seven cells according to a specified form of relationship to age of the individual and for weighting concentration as a function of age in order to determine a hazard function. Thomas[21] remarked that the essence of this model is the peaked weighting function for exposure as a function of age, such that the later the sensitive stage of the model, the later the peak.

Currently, radiation exposures associated with human activity are expected to be low-dose, for example low dose-rate radiation from medical tests, waste cleanup and environmental isolation of materials associated with nuclear weapons and nuclear power production. An exposure-based event can cause a variety of damage scenarios: (1) the damage may be reparable if the damaged cells can repair themselves, and thus there will be no permanent damage; (2) millions of cells may die according to the natural processes of cell death; (3) mutations may occur if the damaged cells exhibit a change in their reproductive structure, resulting in potentially pre-cancerous cells. For such issues, in addition to the frailty model for heterogeneous background presented in this paper, we must consider a model of low-dose exposure based on risk factors describing heterogeneous sensitivity by assuming that each target before the exposure event contains random risk factors. We describe such a model in detail in Appendix B.3.

ACKNOWLEDGEMENTS

We thank Emeritus Professor Hiromitsu Watanabe for providing his unpublished data and Dr. Kenichi Satoh for his suggestions during the development of the models. We also thank the anonymous referees, whose comments were invaluable in the revision of this paper. This work was supported in part by a Grant in Aid (14380123) from the Japanese Ministry of Education, Science and Culture, and by a grant for a Research Program on Low-Dose Radiation Effects Based on Molecular Biology from the

Japan Atomic Energy Research Institute.

(Received November 5, 2004)
(Accepted January 31, 2005)

REFERENCES

1. **Armitage, P. and Doll, R.** 1954. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br. J. Cancer **8:** 1–12.
2. **Dawson, S.V. and Alexeeff, G.V.** 2001. Multistage model estimates of lung cancer risk from exposure to diesel exhaust, based on a U.S. railroad worker cohort. Risk Analysis **21 (No. 1):** 1–18.
3. **Elkind, M.M. and Suton, H.** 1959. X-ray damage and recovery in mammalian cells in culture. Nature **184:** 1293–1295.
4. **Fisher, J.C. and Holloman, J.H.** 1953. A hypothesis for the origin of cancer foci. Cancer **7:** 916–918.
5. **Fujikawa, K., Hasegawa, Y., Matsuzawa, S., Fukunaga, A., Itoh, T. and Kondo, S.** 2000. Dose and dose-rate effects of X rays and fission neutrons on lymphocyte apoptosis in p53 (+/+) and p53 (–/–) mice. J. Radiat. Res. **41:** 113–127.
6. **Hendry, J.H., Potten, C.S., Cadwick, C. and Bianchi, M.** 1982. Cell death (apoptosis) in the mouse small intestine after low doses: effects of dose-rate, 14.7 MeV neutrons, and 600 MeV (maximum energy) neutrons. Int. J. Radiat. Biol. **42:** 611–620.
7. **Hendry, J.H. and Potten, C.S.** 1982. Intestinal cell radiosensitivity: a comparison for cell death assayed by apoptosis or by a loss of clonogenicity. Int. J. Radiat. Biol. **42:** 621–628.
8. **Hougaard, P.** 1984. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. Biometrika **71:** 75–83.
9. **Hougaard, P.** 2000. Analysis of multivariate survival data. Springer-Verlag, New York.
10. **Ibrahim, J., Chen, M.H. and Sinha, D.** 2001. Bayesian survival analysis. Springer-Verlag, New York.
11. **Jensen, S.T., Johansen, S. and Lauritzen, S.L.** 1991. Globally convergent algorithms for maximizing a likelihood function. Biometrika **78:** 867–877.
12. **Kleinbaum, D.G.** 1996. Survival analysis: a self-learning text. Springer-Verlag, New York.
13. **Lehmann, E.L.** 1983. Theory of point estimation. John Wiley & Sons. USA.
14. **Mood, A.M., Graybill, F.A. and Boes, D.C.** 1974. Introduction to the theory of statistics. McGraw-Hill. Singapore.
15. **Moolgavkar, S.H.** 2004. Commentary: Fifty years of the multistage model: remarks on a landmark paper. Int. J. Epidemiol. **33:** 7–8.
16. **Ohara, M., Lu, H., Shiraki, K., Ishimura, Y., Uesaka, T., Katoh, O. and Watanabe, H.** 2001. Radioprotective effects of miso (fermented soy bean paste) against radiation in B6C3F1 mice: increased small intestinal crypt survival, crypt lengths and prolongation of average time to death. Hiroshima J. Med. Sci. **50:** 83–86.
17. **Ohtaki, M., Fujita, S., Hayakawa, N., Kurihara, M. and Munaka, M.** 1985. The age distribution of human adult cancer and an initiation-manifestation model for carcinogenesis. Jpn. J. Clin. Oncol. **15 (Suppl. 1):** 325–343.
18. **Ohtaki, M. and Izumi, S.** 1999. Globally convergent algorithm without derivatives for maximizing a multivariate function. *In* Proceedings of Symposium on "Exploratory Methods and Analyses for Nonlinear Structures of Data with Random Variation" in Hiroshima.
19. **Sahu, S.K. and Dey, D.K.** 2000. A comparison of frailty and other models for bivariate survival data. Lifetime Data Anal. **6:** 207–228.
20. **Takai, A., Kagawa, N. and Fujikawa, K.** 2004. Dose- and time-dependent response for micronucleus induction by x-rays and fast neutrons in gill cells of medaka (*oryzias latipes*). Environ. Mol. Mutagen. **44:** 108–112.
21. **Thomas, D.C.** 1982. Temporal effects and interaction in cancer: Implications of carcinogenic models, p.107–121. *In* R. L. Prentice and A. S. Whittemore (eds.), Environmental epidemiology: Risk assessment, Philadelphia Society for Industrial and Applied Mathematics.
22. **Vaupel, J.W., Manton, K.G. and Stallard, E.** 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography **16:** 439–454.
23. **Whittemore, A.S.** 1977. The age distribution in human cancers for carcinogenic exposures of varying intensity. Am. J. Epidemiol. **106:** 418–432.
24. **Williams, E.D., Lowes, A.P., Williams, D. and Williams G.T.** 1992. A stem cell niche theory of intestinal crypt maintenance based on a study of somatic mutation in colonic mucosa. Am. J. Pathol. **141:** 773–776.

Appendix A

When the exposure dose (D) approaches infinity, the ratio of the survival rate of all targets to the survival rate of one target satisfies the following proposition.

Proposition 1. Let $\rho \geq 1$, and let

$$S_k(D|\beta,\rho) = 1 - \prod_{j=1}^{k}(1 - e^{-\beta\rho^{j-1}D})$$

for $k \geq 1$. Then, it holds that

$$\lim_{D\to+\infty} \frac{S_k(D|\beta,\rho)}{S_1(D|\beta,\rho)} = \begin{cases} k, & \text{if } \rho = 1, \\ 1, & \text{if } \rho > 1. \end{cases}$$

Proof:

If $\rho=1$, then

$$\begin{aligned} \lim_{D\to+\infty} \frac{S_k(D|\beta,\rho)}{S_1(D|\beta,\rho)} &= \lim_{D\to+\infty} \frac{1-(1-e^{-\beta D})^k}{e^{-\beta D}} \\ &= \lim_{D\to+\infty} \frac{k(1-e^{-\beta D})^{k-1}(e^{-\beta D})(-\beta)}{(e^{-\beta D})(-\beta)} \\ &= \lim_{D\to+\infty} k(1-e^{-\beta D})^{k-1} \\ &= k. \end{aligned}$$

If $\rho>1$, then

$$\begin{aligned} \lim_{D\to+\infty} \frac{S_k(D|\beta,\rho)}{S_1(D|\beta,\rho)} &= \lim_{D\to+\infty} \frac{1-\prod_{j=1}^{k}(1-e^{-\beta\rho^{j-1}D})}{e^{-\beta D}} \\ &= \lim_{D\to+\infty} \frac{\sum_{j^*=1}^{k}\left[-\beta\rho^{j^*-1}e^{-\beta\rho^{j^*-1}D}\prod_{j\neq j^*}^{k}(1-e^{-\beta\rho^{j-1}D})\right]}{-\beta e^{-\beta D}} \\ &= \lim_{D\to+\infty} \frac{\sum_{j^*=1}^{k}\left[\rho^{j^*-1}e^{-\beta\rho^{j^*-1}D}\prod_{j\neq j^*}^{k}(1-e^{-\beta\rho^{j-1}D})\right]}{e^{-\beta D}} \\ &= \lim_{D\to+\infty} \left[\prod_{j=2}^{k}(1-e^{-\beta\rho^{j-1}D}) + \sum_{j^*=2}^{k}\left\{\rho^{j^*-1}e^{-\beta(\rho^{j^*-1}-1)D}\prod_{j\neq j^*}^{k}(1-e^{-\beta\rho^{j-1}D})\right\}\right] \\ &= 1. \end{aligned}$$

Appendix B

B.1. Poisson Regression Model

The log-likelihood function of the model as shown in equation (6) is specified as

$$\ell(\theta^*|\boldsymbol{d}_{(obs)}) = \sum_{i=1}^{n} \log P(y_i|D_i,\boldsymbol{x}_i,\theta^*).$$

Then, elements of the Hessian matrix are given by

$$\begin{aligned} \frac{\partial\ell(\theta^*|\boldsymbol{d}_{(obs)})}{\partial\theta_p^*} &= \sum_{i=1}^{n}\left\{\frac{y_i-\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}{\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}\right\}\left[\frac{\partial\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}{\partial\theta_p^*}\right], \\ \frac{\partial^2\ell(\theta^*|\boldsymbol{d}_{(obs)})}{\partial\theta_p^*\,\theta_q^*} &= \sum_{i=1}^{n}\left\{-\frac{y_i\left[\frac{\partial\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}{\partial\theta_p^*}\right]\left[\frac{\partial\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}{\partial\theta_q^*}\right]}{[\mu_k(D_i,\boldsymbol{x}_i|\theta^*)]^2} + \frac{(y_i-\mu_k(D_i,\boldsymbol{x}_i|\theta^*))\left[\frac{\partial^2\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}{\partial\theta_p^*\,\partial\theta_q^*}\right]}{\mu_k(D_i,\boldsymbol{x}_i|\theta^*)}\right\}. \end{aligned}$$

B.2. Gamma-frailty Model for Heterogeneous Background

The log-likelihood function of the model presented in equation (9) is given by

$$\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)}) = \sum_{i=1}^{n} \log f(y_i|D_i, \boldsymbol{x}_i, \boldsymbol{\theta}).$$

Elements of the Hessian matrix are therefore specified as follows:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial \theta_p^*} &= \sum_{i=1}^{n} \frac{(y_i - \mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_p^*}\right]}{\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial \theta_p^* \partial \theta_q^*} &= \sum_{i=1}^{n} \left\{ \frac{\sigma^2\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_p^*}\right]\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_q^*}\right]}{(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))^2} \right. \\
&\quad - \frac{(y_i(1+2\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)))\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_p^*}\right]\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_q^*}\right]}{\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)^2(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))^2} \\
&\quad \left. + \frac{(y_i - \mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))\left[\frac{\partial^2 \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_p^* \partial \theta_q^*}\right]}{\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))} \right\}, \\
\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial \sigma} &= \sum_{i=1}^{n} \left\{ \frac{2\sigma y_i(1+\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))}{(-1+\sigma^2)(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))} + \frac{2}{\sigma^3}\log(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)) - \frac{2\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\sigma^3(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))} \right\}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial \sigma^2} &= \sum_{i=1}^{n} \left\{ \frac{2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*)(3+5\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))}{(\sigma+\sigma^3\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))^2} - \frac{6\log(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))}{\sigma^4} \right. \\
&\quad \left. - \frac{2y_i(1+\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))(1+\sigma^2+\sigma^2(-1+3\sigma^2)\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))}{(-1+\sigma^2)^2(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))^2} \right\}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial \theta_p^* \partial \sigma} &= \sum_{i=1}^{n} \frac{2\sigma(-y_i + \mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))\left[\frac{\partial \mu_k(D_i,\boldsymbol{x}_i|\boldsymbol{\theta}^*)}{\partial \theta_p^*}\right]}{(1+\sigma^2\mu_k(D_i, \boldsymbol{x}_i|\boldsymbol{\theta}^*))^2}.
\end{aligned}$$

In the case of the model for a given exposure dose D, covariates vector $\boldsymbol{x}$=(x_1,x_2, ... ,x_p)^T, and unknown parameters $\boldsymbol{\theta}^* = (k, \mu_0, \beta, \rho, \boldsymbol{\gamma}^T)^T$ having a homogeneous target size expressed as

$$\mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*) = \mu_0 \left\{ 1 - \left(1 - e^{-\beta D e^{\boldsymbol{\gamma}^T \boldsymbol{x}}} \right)^k \right\},$$

we have

$$\begin{aligned}
\frac{\partial \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \mu_0} &= 1 - [F(D|\beta, \boldsymbol{\gamma})]^k, \\
\frac{\partial \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \beta} &= -\frac{\mu_0 D}{\beta} k f(D|\beta, \boldsymbol{\gamma})[F(D|\beta, \boldsymbol{\gamma})]^{k-1}, \\
\frac{\partial \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \gamma_p} &= -x_p \mu_0 D k f(D|\beta, \boldsymbol{\gamma})[F(D|\beta, \boldsymbol{\gamma})]^{k-1}, \\
\frac{\partial \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial k} &= -\mu_0 [F(D|\beta, \boldsymbol{\gamma})]^k \log[F(D|\beta, \boldsymbol{\gamma})], \\
\frac{\partial^2 \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \mu_0^2} &= 0, \\
\frac{\partial^2 \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \mu_0 \partial \beta} &= -\frac{k D}{\beta} f(D|\beta, \boldsymbol{\gamma})[F(D|\beta, \boldsymbol{\gamma})]^{k-1}, \\
\frac{\partial^2 \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \mu_0 \partial \gamma_p} &= -x_p k D f(D|\beta, \boldsymbol{\gamma})[F(D|\beta, \boldsymbol{\gamma})]^{k-1}, \\
\frac{\partial^2 \mu_k(D, \boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \mu_0 \partial k} &= -[F(D|\beta, \boldsymbol{\gamma})]^k \log[F(D|\beta, \boldsymbol{\gamma})],
\end{aligned}$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \beta^2} = -\frac{\mu_0 k D}{\beta^2} h(D|\beta,\gamma)\left\{\frac{k-1}{F(D|\beta,\gamma)} - k\right\} f(D|\beta,\gamma)[F(D|\beta,\gamma)]^{k-1},$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \beta \partial \gamma_p} = -\frac{x_p \mu_0 k D}{\beta}\left\{\frac{D(k f(D|\beta,\gamma) - h(D|\beta,\gamma))}{F(D|\beta,\gamma)} + 1\right\} f(D|\beta,\gamma)[F(D|\beta,\gamma)]^{k-1},$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \beta \partial k} = -\frac{\mu_0 D}{\beta}\{1 + \log[F(D|\beta,\gamma)]\} f(D|\beta,\gamma)[F(D|\beta,\gamma)]^{k-1},$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \gamma_p \partial \gamma_q} = -x_p x_q \mu_0 k\left\{\frac{D(k f(D|\beta,\gamma) - h(D|\beta,\gamma))}{F(D|\beta,\gamma)} + 1\right\} f(D|\beta,\gamma)[F(D|\beta,\gamma)]^{k-1},$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial \gamma_p \partial k} = -x_p \mu_0 D\{1 + k \log[F(D|\beta,\gamma)]\} f(D|\beta,\gamma)[F(D|\beta,\gamma)]^{k-1},$$

$$\frac{\partial^2 \mu_k(D,\boldsymbol{x}|\boldsymbol{\theta}^*)}{\partial k^2} = -\mu_0 [F(D|\beta,\gamma)]^k (\log[F(D|\beta,\gamma)])^2,$$

where $F(D|\beta,\gamma) = 1 - e^{-\beta D e^{\gamma^T \boldsymbol{x}}}$, $f(D|\beta,\gamma) = \beta e^{-\beta D e^{\gamma^T \boldsymbol{x}} + \gamma^T \boldsymbol{x}}$, and $h(D|\beta,\gamma) = \beta e^{\gamma^T \boldsymbol{x}}$.

B.3. Gamma-frailty Model for Heterogeneous Sensitivity (Low Dose)

Let us denote risk factors on the j-th target in each observed individual as z_j, j=1,2, ... , k. For a given exposure dose D, covariates vector $\boldsymbol{x}$=(x_1,x_2, ... ,x_p)^T, and risk factors Z, we construct a model having the form

$$\mu_k(D,\boldsymbol{x}|z,\boldsymbol{\theta}^*) = \mu_0\left\{1 - \prod_{j=1}^{k}(1 - e^{-\beta_j D e^{\gamma^T \boldsymbol{x}} z_j})\right\}. \tag{13}$$

If $\beta_j D$ comes close to zero for $^\forall j$=1,2, ... , k, then the model can be approximately expressed by

$$\mu_k(D,\boldsymbol{x}|z,\boldsymbol{\theta}^*) \simeq \mu_0\left\{1 - \prod_{j=1}^{k}(\beta_j D z_j e^{\gamma^T \boldsymbol{x}})\right\}. \tag{14}$$

By assuming that the sensitivity coefficient of the j-th target (β_j) has regularity following geometrical progression, that is, β_j=$\beta\rho^{j-1}$, the model will be specified by

$$\mu_k(D,\boldsymbol{x}|z,\boldsymbol{\theta}^*) \simeq \mu_0\left\{1 - (\beta\rho^* D \bar{z}^k e^{\gamma^T \boldsymbol{x}})\right\}, \tag{15}$$

where ρ^*=$\rho^{[k(k-1)/2]}$ and $\bar{z}^k$=$\prod_{j=1}^{k} z_j$. Thus, the likelihood function based on the complete data set $\boldsymbol{d}$ is

$$L(\boldsymbol{\theta}^*|\boldsymbol{d}) = \prod_{i=1}^{n} P(y_i|\bar{z}_i, D_i, \boldsymbol{x}_i, \boldsymbol{\theta}^*), \tag{16}$$

where $P(y|\bar{z}, D, \boldsymbol{x}, \boldsymbol{\theta}^*)$ expresses the probability density function of Poisson distribution with mean $\mu_k(D,\boldsymbol{x}|z,\boldsymbol{\theta}^*)$, as shown in equation (15). Integrating the form of the likelihood function in equation (16) with respect to the density function of Z in equation (7) provides the likelihood function based on observed data set $\boldsymbol{d}_{(obs)}$ given by

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)}) &= \int_0^\infty L(\boldsymbol{\theta}^*|\boldsymbol{d})\ \varphi(z_i|\sigma)\ dz_i \\ &= \prod_{i=1}^{n} \int_0^\infty P(y_i|\bar{z}_i, D_i, \boldsymbol{x}_i, \boldsymbol{\theta}^*)\ \varphi(z_i|\sigma)\ dz_i \\ &= \prod_{i=1}^{n} g(y_i|D_i, \boldsymbol{x}_i, \boldsymbol{\theta}). \end{aligned} \tag{17}$$

where

$$g(y|D,\boldsymbol{x},\boldsymbol{\theta}) = \frac{\mu_0^y e^{-\mu_0}}{y!}\left\{\frac{1-(y+\mu_0\sigma)\beta D e^{\boldsymbol{\gamma}^T\boldsymbol{x}}\rho^{\frac{1}{2}k(k-1)}}{\{1-\beta D e^{\boldsymbol{\gamma}^T\boldsymbol{x}}\rho^{\frac{1}{2}k(k-1)}\}^{\frac{1}{\sigma}(1+\sigma)}}\right\}. \tag{18}$$

The log-likelihood function of the model as shown in equation (17) can be specified by

$$\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)}) = \sum_{i=1}^{n}\log g(y_i|D_i,\boldsymbol{x}_i,\boldsymbol{\theta}).$$

Then, elements of the Hessian matrix are

$$\begin{aligned}
\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0} &= \sum_{i=1}^{n}\left\{\frac{y_i}{\mu_0}-1+M(D_i|\boldsymbol{\theta}^*)\Big(\frac{Q(y_i|D_i,\boldsymbol{\theta})}{\mu_0}-\frac{\sigma^2R(y_i|D_i,\boldsymbol{\theta})}{y_i+\mu_0\sigma^2}\Big)\right\},\\
\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\beta} &= \sum_{i=1}^{n}\frac{1}{\beta}T(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\gamma_p} &= \sum_{i=1}^{n}x_pT(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\rho} &= \sum_{i=1}^{n}\frac{k(k-1)}{2\rho}T(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial k} &= \sum_{i=1}^{n}(k-\frac{1}{2})T(y_i|D_i,\boldsymbol{\theta})\log\rho,\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0^2} &= \sum_{i=1}^{n}\left\{\sigma^2M(D_i|\boldsymbol{\theta}^*)U(y_i|D_i,\boldsymbol{\theta})-\frac{y_i}{\mu_0^2}\right\},\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0\partial\beta} &= \sum_{i=1}^{n}\frac{1}{\mu_0}U(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0\partial\gamma_p} &= \sum_{i=1}^{n}x_pU(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0\partial\rho} &= \sum_{i=1}^{n}\frac{k(k-1)}{2\rho}U(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\mu_0\partial k} &= \sum_{i=1}^{n}(k-\frac{1}{2})U(y_i|D_i,\boldsymbol{\theta})\log\rho,\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\beta^2} &= \sum_{i=1}^{n}\frac{(M(D_i|\boldsymbol{\theta}^*))^2}{\beta^2}\left\{\frac{\mu_0\sigma^2Q(y_i|D_i,\boldsymbol{\theta})}{1-\mu_0\sigma^2M(D_i|\boldsymbol{\theta}^*)}-\frac{(y_i+\mu_0\sigma^2)R(y_i|D_i,\boldsymbol{\theta})}{1-(y_i+\mu_0\sigma^2)M(D_i|\boldsymbol{\theta}^*)}\right\},\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\beta\partial\gamma_p} &= \sum_{i=1}^{n}\frac{x_p}{\beta}V(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\beta\partial\rho} &= \sum_{i=1}^{n}\frac{k(k-1)}{2\beta\rho}V(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\beta\partial k} &= \sum_{i=1}^{n}\frac{(2k-1)\log\rho}{2\beta}V(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\gamma_p\partial\gamma_q} &= \sum_{i=1}^{n}x_px_qV(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\gamma_p\partial\rho} &= \sum_{i=1}^{n}\frac{k(k-1)x_p}{2\rho}V(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\gamma_p\partial k} &= \sum_{i=1}^{n}\frac{(2k-1)\log\rho}{2}V(y_i|D_i,\boldsymbol{\theta}),\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\rho^2} &= \sum_{i=1}^{n}\frac{k(k-1)}{4\rho^2}\left\{\Big((k-2)(k+1)+2\mu_0\sigma^2M(D_i|\boldsymbol{\theta}^*)\Big)V(y_i|D_i,\boldsymbol{\theta})-\frac{2y_i[M(D_i|\boldsymbol{\theta}^*)]^2}{1-(y_i+\mu_0\sigma^2)M(D_i|\boldsymbol{\theta}^*)}\right\},\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial\rho\partial k} &= \sum_{i=1}^{n}\frac{k(k-1)(2k-1)}{4\rho^2}\left\{\Big(k(k-1)\log\rho+2(1-\mu_0\sigma^2M(D_i|\boldsymbol{\theta}^*))\Big)V(y_i|D_i,\boldsymbol{\theta})+\frac{2y_i[M(D_i|\boldsymbol{\theta}^*)]^2}{1-(y_i+\mu_0\sigma^2)M(D_i|\boldsymbol{\theta}^*)}\right\},\\
\frac{\partial^2\ell(\boldsymbol{\theta}|\boldsymbol{d}_{(obs)})}{\partial k^2} &= \sum_{i=1}^{n}\frac{\log\rho}{4}\left\{\Big((2k-1)^2\log\rho+4(1-\mu_0\sigma^2M(D_i|\boldsymbol{\theta}^*))\Big)V(y_i|D_i,\boldsymbol{\theta})+\frac{4y_i[M(D_i|\boldsymbol{\theta}^*)]^2}{1-(y_i+\mu_0\sigma^2)M(D_i|\boldsymbol{\theta}^*)}\right\},
\end{aligned}$$

where:

$$
\begin{aligned}
M(D|\theta^*) &= \beta D e^{\gamma^T x} \rho^{\frac{1}{2}k(k-1)}, \\
Q(D|\theta) &= \frac{(1+\sigma^2)\mu_0}{1-\mu_0\sigma^2 M(D|\theta^*)}, \\
R(y|D,\theta) &= \frac{y+\mu_0\sigma^2}{1-(y+\mu_0\sigma^2)M(D|\theta^*)}, \\
T(y|D,\theta) &= M(D|\theta^*)\Big\{Q(D|\theta) - R(y|D,\theta)\Big\}, \\
U(y|D,\theta) &= M(D|\theta^*)\left\{\frac{Q(D|\theta)}{\mu_0(1-\mu_0\sigma^2 M(D|\theta^*))} - \frac{\sigma^2 R(y|D,\theta)}{(y+\mu_0\sigma^2)(1-(y+\mu_0\sigma^2)M(D|\theta^*))}\right\}, \\
V(y|D,\theta) &= M(D|\theta^*)\left\{\frac{Q(D|\theta)}{1-\mu_0\sigma^2 M(D|\theta^*)} - \frac{R(y|D,\theta)}{1-(y+\mu_0\sigma^2)M(D|\theta^*)}\right\}.
\end{aligned}
$$

Appendix C

The algorithm of SPIDER proposed by Ohtaki & Izumi[18] are described as the following steps:

Step 1. Set initial values of the parameters for maximizing of p-dimensional function f, and let denote it as $\alpha_0^{(0)}$.

Step 2. By starting with $\alpha_0^{(s)}$, where s=0,1,2, ... , perform loop at the s stage. Define the function $f_\ell(t) = f(\alpha_{\ell-1}^{(s)} + t\delta_l)$ for ℓ=1, ... , p, where $\delta_\ell = (\delta_{\ell 1}, \delta_{\ell 2}, \cdots, \delta_{\ell p})^T$, a vector of Kronecker's delta. Optimize the function f_ℓ and set

$$
\begin{aligned}
t_\ell &= \arg \max_{t\in(-\infty,+\infty)} f_\ell(t) \\
\alpha_\ell^{(s)} &= \alpha_{\ell-1}^{(s)} + t_q \delta_\ell
\end{aligned}
$$

Step 3. Calculate $\Delta_\ell = \| \alpha_0^{(s)} - \alpha_\ell^{(s)} \|$. If Δ_ℓ becomes small enough, then quit. Otherwise go back to Step 2 with $\alpha_0^{(s+1)} = \alpha_\ell^{(s)}$. Continue Step 2 and Step 3 until convergence.

Japanese Journal of Biometrics Vol. 25, No. 2, 61-67 (2004)

Original Article

Genotyping of Single Nucleotide Polymorphisms Based on a Mathematical Model for Two-Dimensional Data

Kenichi Satoh*[1], Keiko Ohtani*[2], Masaru Ushijima*[3], Minoru Isomura*[3], Masaaki Matsuura*[3], Yoshio Miki*[3] and Megu Ohtaki*[1]

*[1]Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima 734-8553, Japan,
*[2]Japan Biological Informatics Consortium, Hiroshima 734-8553, Japan,
*[3]Genome Center, Japanese Foundation for Cancer Research, Tokyo 170-8455, Japan
e-mail: ohtaki@hiroshima-u.ac.jp

Classification methods typically applied to the Invader assay include k-means clustering and the normal mixture model for original two-dimensional data or angle data. Combining the normal mixture model and angle data might result in an inproved method. In fact, such an approach has the advantages that it can be used to evaluate the goodness of classification for each individual and angle data are easily handled. However, the method requires that the data have an origin, which implies that one cluster must be specified before clustering. Therefore, an alternative method using the normal mixture model is desirable. We propose a mathematical model with a latent time variable. Optimization is based mainly on a one-dimensional normal mixture model with two components, which provides stable computational results more quickly than can be obtained using a bivariate normal mixture model.

Key words: SNP typing; Mathematical model; Normal mixture model.

1. Introduction

Single Nucleotide Polymorphisms (SNPs) are one of the most important biological markers for successfully producing tailor-made medicine (Riva and Kohane, 2002). The Invader assay is one method of genotyping SNPs, in which the resultant two-dimensional data must be classified. Several methods have been considered, such as k-means clustering for original two-dimensional data by Renade *et al.* (2001) , Oliver *et al.* (2002) and van den Oord *et al.* (2003), and the normal mixture model for angle data with a fixed origin by Fujisawa *et al.* (2003). Software is available from the pharmaceutical industry (PerkinElmer Inc., SNPscorer) for aseertaining genotype in the FP-TDI assay; a demonstration version is available

Received July 2004. Revised October 2004. Accepted October 2004.

at http://las.perkinelmer.com/content/snps/software.asp.

We chose to focus on the original two-dimensional data instead of angle data, and propose an alternative clustering method based on a mathematical model that includes a latent time variable and logarithmic transformed data. Optimization is based on a pair of one-dimensional normal mixture models having two components. The approach is rapid and accurate, and is therefore considered suitable for a high-throughput system.

2. Mathematical Model

We assume that there are two types of allele for each SNP, which can be denoted by A and B. Four clusters can be identified in the two-dimensional fluorescence data: 0) there is no DNA, 1) the genotype is homozygous AA, 2) the genotype is homozygous BB, and 3) the genotype is heterozygous AB. Let $\boldsymbol{x} = (x_A, x_B)'$ be an observed vector, where x_j is the fluorescence intensity for allele $j \in \{A, B\}$. Consider the following time-dependent response model:

$$\boldsymbol{x} = (\boldsymbol{v} + \boldsymbol{v}_0)t + \boldsymbol{u}, \tag{1}$$

where

$$\boldsymbol{v} = \begin{pmatrix} y_A z_A \\ y_B z_B \end{pmatrix}, \quad \boldsymbol{v}_0 = \begin{pmatrix} \bar{y}_A y_B e^{\lambda_A} \\ \bar{y}_B y_A e^{\lambda_B} \end{pmatrix} \quad \text{and} \quad \boldsymbol{u} = \begin{pmatrix} \bar{y}_A w_A \\ \bar{y}_B w_B \end{pmatrix}.$$

Here, $y_j \sim Bernoulli(p_j)$ is the latent binary response for allele $j \in \{A, B\}$, $y_j = 1$ if allele j is present, otherwise, $y_j = 0$, $\bar{y}_j = 1 - y_j$, $t \sim LN(0, \delta^2)$ is the latent response time, δ^2 can be zero if there is no correlation between heterozygous clusters, $z_j \sim LN(\mu_j, \tau_j^2)$ is the latent response intensity for the case in which allele j is present, $w_j \sim LN(\nu_j, \sigma_j^2)$ is the latent response intensity for the other cases, $\mu > \nu$, $\{y, t, z, w\}$ are mutually independent, and λ_j indicates the background (apparent) response that accounts for the gradient from a homozygous cluster to the heterozygous cluster.

We now turn attention to the parameter λ_j. From Model (1), data belonging to clusters 0 and 2 can be expressed as $(x_A, x_B) = (w_A, w_B)$ and $(w_A + e^{\lambda_A} t, z_B t)$, respectively. The difference between means on the x_A-axis is $e^{\lambda_A + \delta^2/2}$, which is empirically small, so restrict attention to the special case

$$\boldsymbol{v}_0 = 0. \tag{2}$$

Under condition (2), the logarithmic transformed observations, $\tilde{x}_j = \ln(x_j)$ and $\tilde{\boldsymbol{x}} = (\tilde{x}_A, \tilde{x}_B)'$, have the following marginal and joint distribution:

$$f_j(\tilde{x}_j) = p_j \phi^{(1)}(\tilde{x}_j \mid \mu_j, \tau_j^2 + \delta^2) + q_j \phi^{(1)}(\tilde{x}_j \mid \nu_j, \sigma_j^2) \quad \text{for} \quad j \in \{A, B\} \quad \text{and} \quad f(\tilde{\boldsymbol{x}}) = \sum_{l=0}^{3} \xi_l g_l(\tilde{\boldsymbol{x}}) \tag{3}$$

where

$$\begin{aligned}
\xi_0 &= q_A q_B, \quad g_0(\tilde{\boldsymbol{x}}) = \phi^{(1)}(\tilde{x}_A \mid \nu_A, \sigma_A^2)\phi^{(1)}(\tilde{x}_B \mid \nu_B, \sigma_B^2),\\
\xi_1 &= p_A q_B, \quad g_1(\tilde{\boldsymbol{x}}) = \phi^{(1)}(\tilde{x}_A \mid \mu_A, \tau_A^2 + \delta^2)\phi^{(1)}(\tilde{x}_B \mid \nu_B, \sigma_B^2),\\
\xi_2 &= q_A p_B, \quad g_2(\tilde{\boldsymbol{x}}) = \phi^{(1)}(\tilde{x}_A \mid \nu_B, \sigma_A^2)\phi^{(1)}(\tilde{x}_B \mid \mu_B, \tau_B^2 + \delta^2),\\
\xi_3 &= p_A p_B, \quad g_3(\tilde{\boldsymbol{x}}) = \phi^{(2)}(\tilde{\boldsymbol{x}} \mid \boldsymbol{\mu}, \Sigma),
\end{aligned}$$

with $p_j + q_j = 1$, $p_j \geq 0$, $q_j \geq 0$, $\boldsymbol{\mu} = (\mu_A, \mu_B)'$, $\Sigma = diag(\tau_A^2, \tau_B^2) + \delta^2 \mathbf{1}\mathbf{1}'$, and $\phi^{(k)}(\cdot \mid \boldsymbol{\theta}, \Omega)$ is the probability density function of the k-variate normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix Ω. The transformed data can therefore be expressed as a normal mixture model (see McLachlan and Peel, 2000) under condition (2). Note that the homozygous cluster consists of independent variables under Model (3).

3. Estimation and Clustering

Suppose that the parameters of Model (3) are known. The observation can be clustered using Bayes rule (Wolfe, 1970). Let $l = l_A + 2 \cdot l_B$ for $l_j \in \{0,1\}$ and $j \in \{A,B\}$ and w_l be the posterior probability that $\tilde{\boldsymbol{x}}$ belongs to Cluster l for $l = 0, \cdots, 3$, which is given by the following equation:

$$w_l(\tilde{\boldsymbol{x}}) = Pr(y_A = l_A \quad \text{and} \quad y_B = l_B \mid \tilde{\boldsymbol{x}}) = \frac{\xi_l g_l(\tilde{\boldsymbol{x}})}{f(\tilde{\boldsymbol{x}})}. \tag{4}$$

An individual is classified into cluster l_0, where $l_0 = \arg_{l=0,\cdots,3} \max w_l(\tilde{\boldsymbol{x}})$. The average of the maxima of these posterior probabilities is known as an allocation rate, which assesses the performance of the mixture approach to clustering (e.g. Basford and McLachlan (1985)).

When the parameters of Model (3) are unknown, they can be estimated by the following two steps: 1) estimate each marginal distribution using the EM algorithm (Dempster and Laird *et al.*, 1977; McLachlan and Krishnan, 1997), and 2) estimate the joint distribution and the unknown covariance δ.

Unfortunately, the clustering rule, as based on the posterior probability, can at times provide unreasonable clusters because the variance of Cluster 0 is much smaller than that of Cluster 1 or Cluster 2 in the marginal distribution, and a part of Cluster 0 might be mis-classified as a homozygous cluster. To overcome this difficulty, we consider a shift-logarithmic transformation, so that variances of the marginal distributions can be made equivalent in the following manner: 1) for a given value α, transform the data by $\tilde{x}^{(\alpha)} = \ln(x - \alpha)$, 2) fit the marginal distribution to the transformed data and estimate the variances $\widehat{\tau^2 + \delta^2}^{(\alpha)}$ and $\widehat{\sigma^2}^{(\alpha)}$, 3) repeat operations 1) and 2) until the optimized shift value α^* satisfies $\widehat{\tau_j^2 + \delta^2}^{(\alpha^*)} = \widehat{\sigma^2}^{(\alpha^*)}$, 4) fit the joint distribution to the transformed data $\tilde{\boldsymbol{x}}^{(\alpha^*)}$ where $\boldsymbol{\alpha}^* = (\alpha_A^*, \alpha_B^*)'$ and estimate the covariance, and 5) identify four clusters according to the fitted joint distribution. The existence of the optimized shift value α^* depends on the minimum value of the observations. Thus α^* can be formally written as follows:

$$\alpha^* = \frac{rE(w) - E(zt)}{r - 1}, \quad \text{where} \quad r^2 = \frac{Var(zt)}{Var(w)}.$$

Note that the estimates of the mixing proportions p_j, $j \in \{A, B\}$, might be close to one or zero, which implies that the potential total number of clusters is one or two. In such a case, the validity of the clustering result needs to be assessed not only statistically but also biologically.

4. Examples and Discussion

The Invader assay generates two-dimensional fluorescence intensities. We examined the performance of our method using three examples of SNP data, which were selected from a large number of SNPs available from the Japanese Foundation for Cancer Research (JFCR). All were obtained with informed consent.

Original two-dimensional data are plotted on the left side in Figures 1-3. The fluorescence intensity for allele A (x_A) is plotted on the horizontal axis, that for allele B (x_B) on the vertical axis. Figure 1 shows ideal data, in which the observed results are clearly split, response time (or rate) is thus sufficiently long, and response intensities appear to be saturated. On the other hand, the response time of the data in Figure 2 is shorter and the boundaries among clusters are

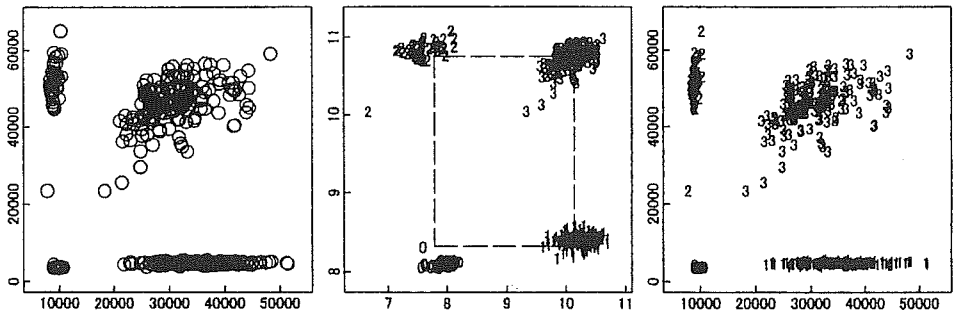

Fig. 1. Example data showing a long response time. Left: scatter plot of original fluorescence intensities. Center: logarithmic transformed observations and clustering results. Right: restored results. The allocation rate of the clustering result is 100.0%. Crosses represent individuals with low posterior probability.

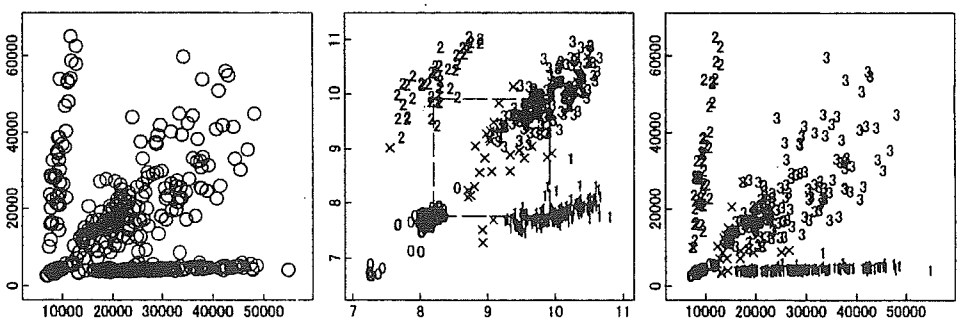

Fig. 2. Example data showing a short response time. Left: scatter plot of original fluorescence intensities. Center: logarithmic transformed observations and clustering results. Right: restored results. The allocation rate of the clustering result is 98.5%. Crosses represent individuals with low posterior probability.

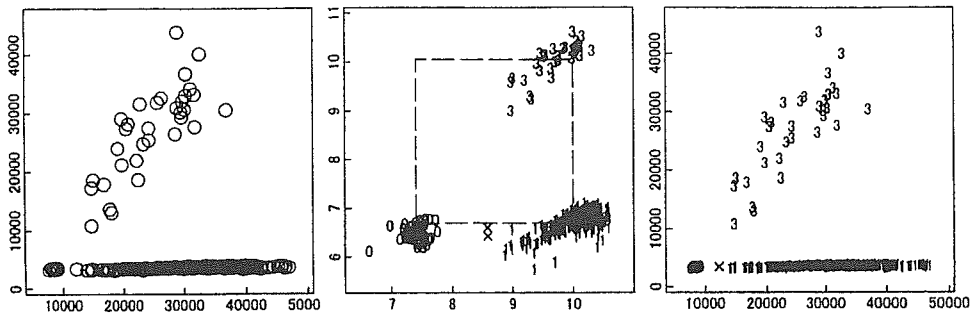

Fig. 3. Example data showing only three clusters. Left: scatter plot of original fluorescence intensities. Center: logarithmic transformed observations and clustering results. Right: restored results. The allocation rate of the clustering result is 99.9%. Crosses represent individuals with low posterior probability.

unclear. Figure 3 illustrates a case in which Cluster 2 is not observed, so there are only three clusters.

For each original data set, the optimized shift value α and its covariance matrix ware estimated. The logarithmic transformed data and original-scale data were then plotted at the center and right of each Figure, respectively. The plotted number expresses the cluster number determined by Bayes rule, and the crosses represent individuals with low posterior probability. The four estimated means are connected by a dotted line. Nine further examples of clustering results are shown in Figure 4. The allocation rate value for each clustering result is superimposed. The individuals with low posterior probability are not distinguished in Figure 4.

Our clustering method is based on Model (3) instead of Model (1). Therefore the estimators of the means, such as the mean of Cluster 0 at the center of Figure 1, are not unbiased. However, we make use of the following four advantages: 1) the joint distribution is expressed in an explicit form; 2) estimators of the marginal distributions can be used for estimation of the joint distribution, such that numerical calculations are easier and faster; 3) regardless of whether or not the real response time is long, four clusters can be properly identified because estimation is primarily conducted on the marginal distributions; and 4) even when there are only three clusters, the joint distribution can be estimated due to the mean structure of the rectangle.

If Condition (2) does not hold, cluster 1 or 2 will be shifted towards the inside of the rectangle, the diagonal of which is defined by clusters 0 and 3 (see the vertical axis of the left top example in Figure 4). In such cases, the marginal distribution might have three components so our proposed method based on two components will not work efficiently. Therefore testing or selecting the number of components on the marginal distribution is useful, as described , for example, by Nakamura and Konishi (1998).

Although our purpose is to classify individuals properly, better estimators of the means and covariance matrices in the normal mixture model might be required for drawing the contour of

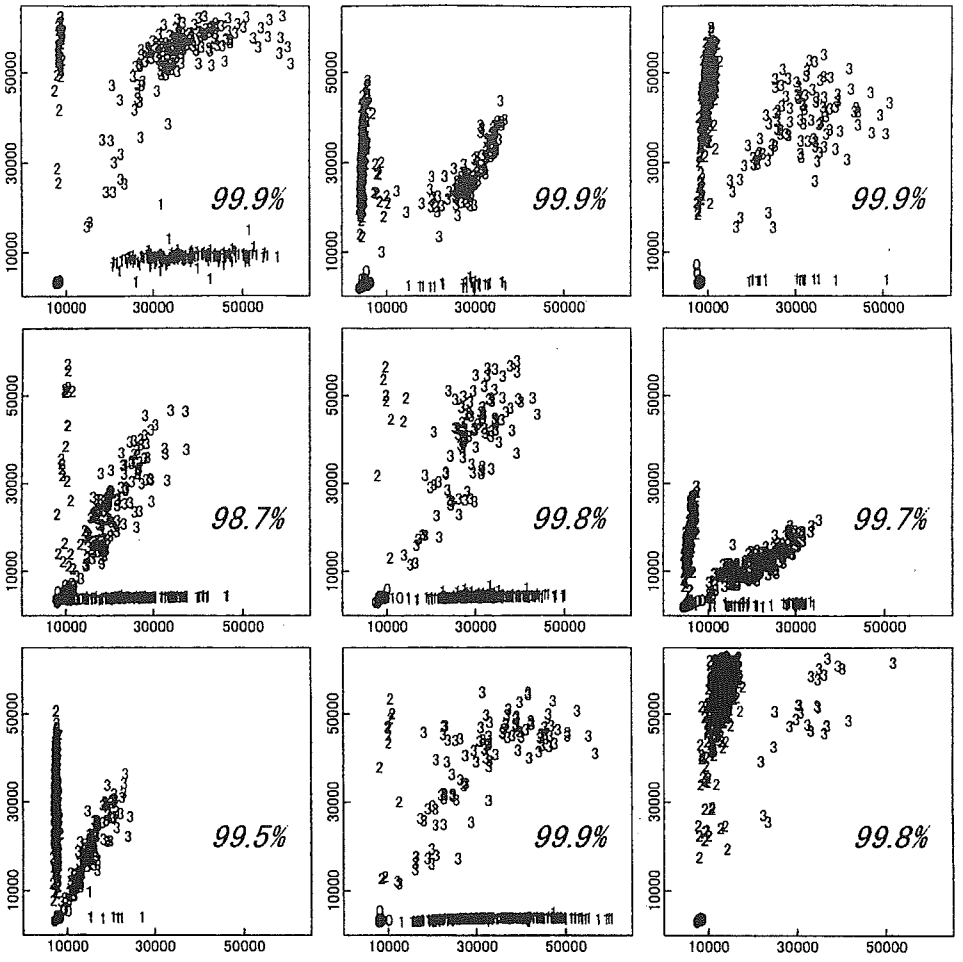

Fig. 4. Nine further examples of clustering results. The allocation rate value for each clustering result is superimposed. Note that individuals with low posterior probability are not distinguished.

the probability density function or finding outliers located in the tail. If the allocation rate is close to one, weighted means and covariance matrices derived from posterior probabilities are useful, which provide almost the same clustering result. Otherwise, the clustering result of model (3) itself might not be reliable and we need to check the goodness of the model fit, the number of potential clusters in the observation and the validity of the experiment.

REFERENCES

Basford, K. E. and McLachlan, G. J. (1985). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association* **80**, 286-293.

Dempster, A. P., Laird, N, M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B **39**, 1-38.

Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto, T. and Matsuura, M. (2003). Genotyping of single nucleotide polymorphism using model-based clustering. *Bioinformatics*, **20**, 718 - 726.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Nakamura, N. and Konishi, S. (1998). Estimation of number of components for multivariate normal mixture models based on information criteria. *Japanese Journal of Applied Statistics*, **27**, 165-180, in Japanese.

Oliver, M., Chuang, L.-M., Chang, M.-S., Chen, Y.-T., Pei, D., Ranade, K., de Witte, A., Allen, J., Tran, N., Curb, D., Pratt, R., Neefs, H., de Arruda Indig, M., Law, S., Neri, B., Wang, L. and Cox, D. R. (2002). High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Res.* **30**, e53.

Ranade, K., Chang, M.-S., Ting, C.-T., Pei, D., Hsiao, C.-F., Oliver, M., Pesich, R., Hebert, J., Chen, Y.-D., Dzau, V. J., Curb, D., Olshen, R., Risch, N., Cox, D. R., and Botstein, D. (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Res.* **11**, 1262-1268.

Riva, A. and Kohane, I. S. (2002). SNPper: retrieval and analysis of human SNPs. *Bioinformatics* **18**, 1681-1685.

van den Oord, E. J. C. G., Jiang, Y., Riley, B. P., Kendler, K. S. and Chen X. (2003). FP-TDI SNP Scoring by Manual and Statistical Procedures: A Study of Error Rates and Types. *BioTechniques* **34**, 610-624.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Res.* **5**, 329-350.

Japanese Journal of Biometrics Vol. 26, No. 1, 31-48 (2005)

Preliminary Report

Model-based analysis of microarray data: Exploration of differentially expressed genes between two cell types based on a two-dimensional mixed normal model

Megu Ohtaki[*1,†], Keiko Otani[*2], Kenichi Satoh[*1], Toshihiko Kawamura[*3], Keiko Hiyama[*4] and Masahiko Nishiyama[*4]

[*1]Department of Environmetrics and Biometrics,
Research Institute for Radiation Biology and Medicine,
Hiroshima University, Hiroshima, Japan
[*2]Japan Biological Informatics Consortium
[*3]Hiroshima Cancer Therapy Development organization
[*4]Department of Translational Cancer Research,
Research Institute for Radiation Biology and Medicine,
Hiroshima University, Hiroshima, Japan
[†]e-mail: ohtaki@hiroshima-u.ac.jp

Inference on gene expression change between two different samples is considered. We develop a mathematical model assuming that there exist two different functional states of a gene: "ON" and "OFF" . Each measured sample-specific gene expression intensity is described by an additive model, which accounts for fluctuations in absolute gene expression intensity and measurement error, to which a two-dimensional mixed normal model with four components considering the joint distribution of the sample "sum" and "difference" is approximated. We can successfully identify genes that are differentially expressed between two samples using posterior probabilities, while avoiding declaring false differences. The proposed methods are applicable to cDNA microarray data with two fluorescent dyes and to oligonucleotide data.

Key words: cDNA microarray; Empirical Bayes; Gene Expression; Mixed Normal Distribution; Normalization; Oligonucleotide microarray

1. Introduction

DNA microarray technology is presently the most effective high-throughput tool for identifying specific genes among tens of thousands of background genes (Gerhold et al., 2002; Schena et al., 1995). In cDNA microarray experiments, genes differentially expressed between query and reference samples (e.g. cancer and noncancerous cells) are identified through many processes using two different fluorescent dyes. The microarray experimental procedure proceeds as follows. **RNA isolation:** Two mRNA samples to be compared are isolated from the query and reference samples and reverse transcribed into cDNA.

Received December 2004. Revised May 2005. Accepted June 2005.

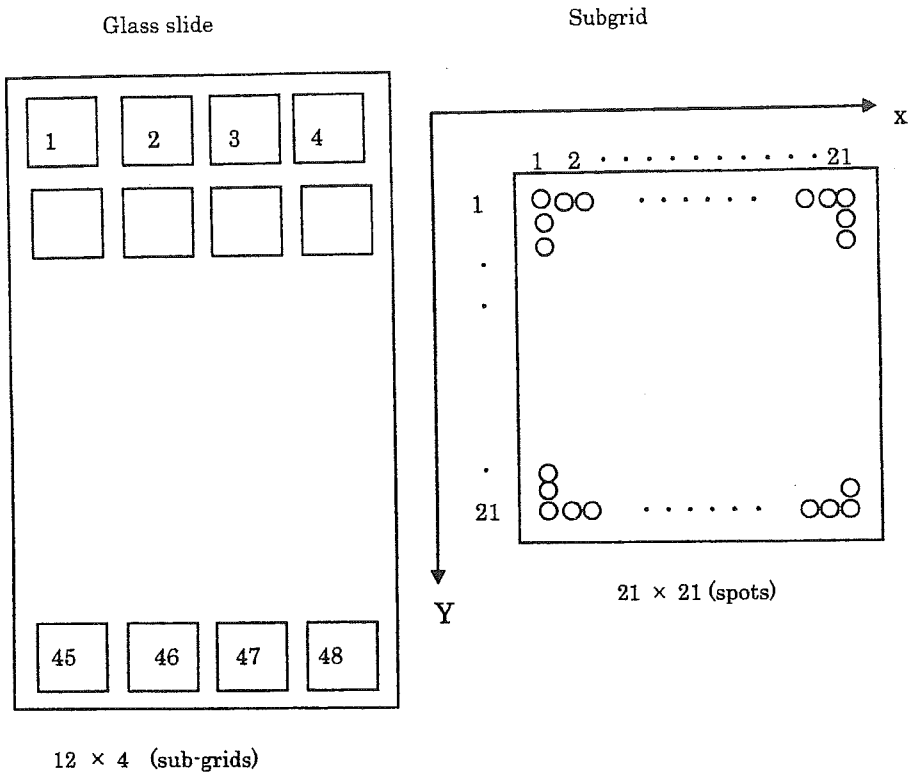

Fig. 1. A diagram of the microarray glass slide is shown schematically. It contains 48 subgrids, each containing 441 spots. In total, 21168 genes are spotted in a microarray.

Sample labeling: The cDNA from each sample is labeled using either a red fluorescent dye (cy5) or a green fluorescent dye (cy3).

Hybridization: Equal quantities of the two differentially labeled samples are mixed and hybridized to a microarray containing 21168 cDNA probes.

Array preparation: Array preparation depends on the experimental design. There are many variations in the geometrical setup of the microarray glass slide and gridding head used for spotting. We used a glass slide, Riken human 21K array, with a subgrid of 48 blocks. Each block contained 441 spots (21 × 21). Each spot on the slide is numbered from 1 to 21168 (48 × 441), which we call gene IDs. An (i, j) coordinate indicates the location of a spot at the i-th row and j-th column in a subgrid. The midpoint of a subgrid is denoted by (m_r, m_c). Figure 1 shows a schematic diagram of the microarray slide.

Data collection: The slides are imaged using a scanner and fluorescence measurements are made separately at each spot on the array by channels 1 and 2 for the two dyes. The data obtained from each channel consists of foreground and background intensities.

The purpose of our study is to successfully identify differentially expressed genes between the two samples, while avoiding false positives (i.e. declaring unchanged genes as differentially expressed). We propose a new mathematical model for microarray data based on a hypothesis

for functional status of the genes, and a method for estimating the probability of a gene being expressed differentially in the two samples being compared.

In Section 2, we introduce the mathematical model for microarray data. So far several model-based approaches have been proposed for identification of differentially expressed genes. They are divided roughly into the following two methods: ANOVA based on the fixed effects linear model (Churchill et al., 2002; Kerr et al., 2000; Lee et al., 2000) and the empirical Bayes method (Baldi et al., 2001; Kendziorski et al., 2003; Long et al., 2001; Newton et al., 2001). An empirical Bayes approach that treats the gene expression intensities as arising from some population was originally proposed by Newton et al. (2001). We also employ the two-group empirical Bayes method to infer differentially expressed genes between two samples. The point in which our approach differs from theirs is that the functional status of a gene is introduced into the mathematical model using a binary variable. We assume there exist two different functional states of a gene: "ON" and "OFF". Biologically, "ON" means the gene produces mRNA and "OFF" means the gene does not produce it. In cases where a gene is "ON", the gene expression intensity is regarded as the "sum" of a random variable that obeys the log normal distribution and a measurement error. In cases where a gene is "OFF", the gene expression intensity is regarded as measurement error only.

We use the "sum" and "difference" simultaneously, for which the variable transformation named S-D transform is introduced in Section 2.1. The acronym "S-D" is used for an abbreviation of "sum" and "difference" of gene expression levels in the query and reference samples. Here the scatter plot using these variables is called an S-D plot. In Section 2.2, a two-dimensional mixed normal density function having at most four components is introduced as the joint distribution of the S-D transformed variables. Section 2.3 describes the exploration of differentially expressed genes. The probability of gene i expressing differently between query and reference samples is obtained as a posterior probability. In Section 3, we explain the implementation of our method to detect differentially expressed genes using real cDNA microarray data. Because a massive amount of data is generated by cDNA microarray experiments, there could be large experimental variations that affect the resultant estimated gene expression levels. Many researchers stress the importance of normalization before carrying out the statistical analysis (Dudoi et al., 2002; Fan et al., 2004; Saviozzi et al., 2003; Schudhhardt et al., 2002; Yang et al., 2002; Wu et al., 2001). We describe a procedure of normalization that is based directly on the mathematical model in Section 3.1. Section 3.2 describes the parameter-estimation procedure.

Though we focus on cDNA microarrays in the present development, the proposed model is also applicable for analyzing a pair of expression data from oligonucleitide arrays (Irizarry et al., 2003; Li et al., 2001).

2. Mathematical model for microarray data

Suitably transformed (the logarithmic transformation in this study) and normalized gene expression measurements are expressed with a simple additive mathematical model that accounts for measurement error and fluctuations in absolute gene expression levels for each channel. We consider two types of measurement error in microarray data. One is common to channels 1 and 2, the other is detected independently between the two channels. Let $Y_i^{(1)}$ and $Y_i^{(2)}$ be the expression intensities of gene g_i in the query and reference samples, respectively, which are suitably transformed and normalized. The mathematical model for microarray data can be described as follows:

$$\begin{cases} Y_i^{(1)} = \tau_i^{(1)}\alpha_i\rho_i + \beta_i + \varepsilon_i^{(1)}, \\ Y_i^{(2)} = \tau_i^{(2)}\alpha_i\rho_i + \beta_i + \varepsilon_i^{(2)}. \end{cases} \tag{1}$$

The symbols $\tau_i^{(1)}$ and $\tau_i^{(2)}$ represent the expression status of gene g_i in the query and reference samples, respectively, which are defined by

$$\tau_i = \begin{cases} 1 & \text{if } g_i \text{ is "ON"}, \\ 0 & \text{if } g_i \text{ is "OFF"}. \end{cases}$$

The symbol α_i represents the expression level of gene g_i when it is "ON". The symbol ρ_i represents the volume of cDNA probe, which includes the fluctuation in probe volume. Since α_i and ρ_i are not identifiable unless repeated measurement are available, we replace $\alpha_i\rho_i$ by α_i for simplicity. We regard it as a positive random variable having log-normal distribution with mean $\log\mu - \frac{\lambda^2}{2}$ and variance λ^2 (*i.e.* $\log\alpha_i \sim N(\log\mu - \frac{\lambda^2}{2}, \lambda^2)$). Thus, $E(\alpha_i) = \mu$ and $Var(\alpha_i) = \mu^2(e^{\lambda^2} - 1)$. The symbols β_i denote random errors common to channels 1 and 2, which obey a normal probability density function with mean 0 and variance σ_β^2 (i.e. $\beta_i \sim i.i.d.\ N(0, \sigma_\beta^2)$). With oligonucleotide maicroarrays, each microarray measures a single sample and provides an absolute measurement level for each RNA molecule (Butte, 2002). Therefore the term β_i should be negligible in the case of oligonucleotide microarray. The symbols $\varepsilon_i^{(1)}$ and $\varepsilon_i^{(2)}$ indicate random errors, which are mutually independent and have normal probability density functions with mean zero and variance σ_ε^2 (i.e. $\varepsilon_i^{(1)}, \varepsilon_i^{(2)} \sim i.i.d.\ N(0, \sigma_\varepsilon^2)$).

2.1 S-D transformation

The S-D transformation of paired expression intensities ($Y_i^{(1)}$, $Y_i^{(2)}$) into the "sum" and "difference". (U_i, V_i) can be described mathematically as follows:

$$\begin{cases} U_i = Y_i^{(1)} + Y_i^{(2)} = (\tau_i^{(1)} + \tau_i^{(2)})\alpha_i + 2\beta_i + \varepsilon_i^{(1)} + \varepsilon_i^{(2)}, \\ V_i = Y_i^{(1)} - Y_i^{(2)} = (\tau_i^{(1)} - \tau_i^{(2)})\alpha_i + \varepsilon_i^{(1)} - \varepsilon_i^{(2)}. \end{cases} \tag{2}$$

Then the conditional mean and variance of U and V are as follows:

$$E[U|(\tau^{(1)}, \tau^{(2)})] = \begin{cases} 0, & (\tau^{(1)}, \tau^{(2)}) = (0, 0), \\ 2\mu, & (\tau^{(1)}, \tau^{(2)}) = (1, 1), \\ \mu, & (\tau^{(1)}, \tau^{(2)}) = (1, 0), \\ \mu, & (\tau^{(1)}, \tau^{(2)}) = (0, 1), \end{cases} \tag{3}$$

$$Var[U|(\tau^{(1)}, \tau^{(2)})] = \begin{cases} 4\sigma_\beta^2 + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (0, 0), \\ 4\mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (1, 1), \\ \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (1, 0), \\ \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (0, 1), \end{cases} \tag{4}$$

$$E[V|(\tau^{(1)}, \tau^{(2)})] = \begin{cases} 0, & (\tau^{(1)}, \tau^{(2)}) = (0, 0), \\ 0, & (\tau^{(1)}, \tau^{(2)}) = (1, 1), \\ \mu, & (\tau^{(1)}, \tau^{(2)}) = (1, 0), \\ -\mu, & (\tau^{(1)}, \tau^{(2)}) = (0, 1), \end{cases} \tag{5}$$

$$Var[V|(\tau^{(1)}, \tau^{(2)})] = \begin{cases} 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (0, 0), \\ 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (1, 1), \\ \mu^2(e^{\lambda^2} - 1) + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (1, 0), \\ \mu^2(e^{\lambda^2} - 1) + 2\sigma_\varepsilon^2, & (\tau^{(1)}, \tau^{(2)}) = (0, 1). \end{cases} \tag{6}$$

The concept of the S-D transformation for the four possible states of $(\tau_i^{(1)}, \tau_i^{(2)})$ is visually illustrated in Figure 2. A gene inside region A is "OFF" in both samples, a gene inside region B is "ON" in both samples, and a gene in region C or D is "ON" in one sample and "OFF" in the other.

Under the condition that $\tau_i^{(1)} = \tau_i^{(2)}$ (i.e. gene g_i is "ON" or "OFF" in both samples), the variable V follows the normal distribution $N(0, 2\sigma_\varepsilon^2)$, in which the value V_i shows only measurement error. Then the equations

$$E(V|U = u) = 0, \tag{7}$$

$$Var(V|U = u) = 2\sigma_\varepsilon^2, \tag{8}$$

hold regardless of the value U.

Using the inverse S-D transformation, we obtain the equations

$$\begin{cases} Y_i^{(1)} = \frac{1}{2}(U_i + V_i), \\ Y_i^{(2)} = \frac{1}{2}(U_i - V_i). \end{cases} \tag{9}$$

2.2 Two-dimensional mixed normal model

We introduce a two-dimensional normal mixture model having at most four components as the joint distribution of (U, V), whose density function is given by

$$f(u, v|\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{s,t \in \{0,1\}} p_{st} f_{st}(u, v|\boldsymbol{\theta}), \tag{10}$$