Improved Parameter Estimation for Variance-Stabilizing Transformation of Gene-Expression Microarray Data

MASATO INOUE

Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute,
Saitama 351-0198, Japan
"Intelligent Cooperation and Control", PRESTO, Japan Science and Technology Corporation,
c/o RIKEN BSI, Saitama 351-0198, Japan
Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan
E-mail: minoue@brain.riken.jp

SHIN-ICHI NISHIMURA

Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute

Graduate School of Medicine, University of Tokyo, Tokyo 113-0033, Japan

GEN HORI

Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute

HIROYUKI NAKAHARA

Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute

MICHIKO SAITO and YOSHIHIRO YOSHIHARA

Laboratory for Neurobiology of Synapse, RIKEN Brain Science Institute

SHUN-ICHI AMARI

Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute

A gene-expression microarray datum is modeled as an exponential expression signal (log-normal distribution) and additive noise. Variance-stabilizing transformation based on this model is useful for improving the uniformity of variance, which is often assumed for conventional statistical analysis methods. However, the existing method of estimating transformation parameters may not be perfect because of poor management of outliers. By employing an information normalization technique, we have developed an improved parameter estimation method, which enables statistically more straightforward outlier exclusion and works well even in the case of small sample size. Validation of this method with experimental data has suggested that it is superior to the conventional method.

Keywords: gene-expression microarray data; variance stabilization; information normalization.

2   *M. Inoue et al.*

1. Introduction

A typical use of gene-expression microarray data is to detect differentially expressed genes under relevant biological conditions. For this purpose, analysis of variance (ANOVA) is applicable, but its assumption of uniform variance is often violated by microarray data, resulting in unreliable p-values. This characteristic of microarray data can be explained by a model of an exponential signal and additive noise.[1] Moreover, variance-stabilizing transformation based on this model can improve variance uniformity, leading to more accurate and sensitive statistical results.[2,3,4,5]

Problems arise owing to the assumptions underlying analysis methods. If we had 'exact' analysis methods that could deal with any given distribution, we would not need variance-stabilizing transformation; in fact, variance stabilization does not affect the results of 'exact' analyses. However, such ideal analysis methods are difficult to realize. Therefore, many conventional analysis methods employ the assumptions of uniform variance and normality. Even non-parametric tests are not always an exception to this; e.g., a non-parametric version of ANOVA, the Kruskal-Wallis test, assumes identical distributions (but allows shifting) for each group, thus uniform variances are indispensable. What we can do here is apply variance stabilization to data before using conventional analysis methods.

Variance-stabilizing transformation,[6] in brief, modifies the unit or the metric of given data to a statistically desirable one. All given data are transformed by a common function, which is a monotonically increasing one-to-one correspondence function. What actually enables variance stabilization is determination of the dependence of each variance on the average for the given data. In the case of gene-expression microarray data, based on the above model, this dependence is a quadratic function consisting of three parameters: the coefficient of signal variation, the noise variance, and the noise mean. This transformation would be useful in most cases unless we were interested in these three estimators themselves. The superiority of this transformation compared to other alternative methods such as the 'started logarithm' has been suggested.[7]

In this paper, we propose a new method for estimating these three parameters because the conventional method[4,5] is disturbed by outliers. In the model section, we define the gene-expression microarray data model and summarize the general framework of variance-stabilizing transformation. In the theory section following that, we describe the conventional parameter estimation, and then propose our new method. The effectiveness of our method is compared to that of the conventional method based on experimental data in the validation section. Then we discuss the proposed method from another statistical point of view.

2. Model

Here we define the stochastic model of gene-expression microarray data and form a variance-stabilizing transformation based on this model. The datum of the ith gene under the jth condition is defined as the stochastic variable; that is, the summation of a signal with log-normal distribution and random noise Q with an unknown but common distribution to

all genes and conditions,

$$X_{ij} = \exp Z_{ij} + Q, \tag{1}$$

where Z_{ij} is taken from a normal distribution with mean μ_{ij} and variance σ_{ij}^2 (we use the expression of $\mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$). All Z_{ij} and Q are independent of each other. We sometimes use $X_{ij}^{(k)}$, which denotes the kth stochastic variable among N_{ij} observations, and its sample value $x_{ij}^{(k)}$. Usually, N_{ij} is a small value, such as three or four, because of the high cost of a microarray experiment. We assume each σ_{ij}^2 is equal to σ_Z^2. We also assume all data have been well calibrated between each microarray chip.

   We notice that every variance of X_{ij} can be determined using the corresponding average of X_{ij} through a common quadratic function $f_{\boldsymbol{\theta}}$,

$$\mathrm{Var}[X_{ij}] = f_{\boldsymbol{\theta}}(\mathrm{E}[X_{ij}]), \tag{2}$$

$$f_{\boldsymbol{\theta}}(t) \equiv (e^{\sigma_Z^2} - 1)(t - \mu_Q)^2 + \sigma_Q^2, \tag{3}$$

where $\mu_Q \equiv \mathrm{E}[Q]$, $\sigma_Q^2 \equiv \mathrm{Var}[Q]$, and $\boldsymbol{\theta} \equiv \{\sigma_Z^2, \mu_Q, \sigma_Q^2\}$, while $\mathrm{E}[\bullet]$ and $\mathrm{Var}[\bullet]$ denote the average and variance, respectively. Generally, if we know f, the function of the variance dependence on its average, we can form the transformation function g that stabilizes variances to 1 using the first-order Taylor approximation,[6]

$$
\begin{aligned}
1 &= \mathrm{Var}[\, g(X) \,] \\
&\simeq \mathrm{Var}[\, g(\mathrm{E}[X]) + g'(\mathrm{E}[X])(X - \mathrm{E}[X]) \,] \\
&= \{g'(\mathrm{E}[X])\}^2 f(\mathrm{E}[X]),
\end{aligned}
\tag{4}
$$

$$g(x) = \int^x \frac{1}{\sqrt{f(t)}} dt, \tag{5}$$

where g' is the first derivative of g.

   Here we use a modified $f_{\boldsymbol{\theta}}$ so that the corresponding transformation $g_{\boldsymbol{\theta}}$ conforms to the conventional log transformation under noiseless circumstances ($\mu_Q = \sigma_Q^2 = 0$),

$$f_{\boldsymbol{\theta}}(t) \equiv \sigma_Z^2(t - \mu_Q)^2 + \sigma_Q^2, \tag{6}$$

$$g_{\boldsymbol{\theta}}(x) = \frac{1}{\sigma_Z} \ln h_{\boldsymbol{\theta}}(x), \tag{7}$$

$$h_{\boldsymbol{\theta}}(x) \equiv \frac{1}{2} \left(\sqrt{(x - \mu_Q)^2 + \sigma_Q^2/\sigma_Z^2} + x - \mu_Q \right), \tag{8}$$

where we have used $e^{\sigma_Z^2} - 1 \simeq \sigma_Z^2$. This transformation function $g_{\boldsymbol{\theta}}$ is a monotonically increasing function, hence it guarantees order preservation and one-to-one correspondence. The function $h_{\boldsymbol{\theta}}$ can be used as a noise reduction function. Now, what we need to know is the transformation parameters $\boldsymbol{\theta}$.

4   *M. Inoue et al.*

3. Theory

In this section, we briefly summarize the conventional estimation method and propose a more straightforward and potentially better one. Before we get onto our main subject, we do several definitions using sample mean $M_{ij} \equiv \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} X_{ij}^{(k)}$ and unbiased variance $V_{ij} \equiv \frac{1}{N_{ij}-1} \sum_{k=1}^{N_{ij}} (X_{ij}^{(k)} - M_{ij})^2$. We also use the transformed variable $\tilde{X}_{ij}^{(k)} \equiv g_{\boldsymbol{\theta}}(X_{ij}^{(k)})$, corresponding to sample mean $\tilde{M}_{ij}$ and unbiased variance $\tilde{V}_{ij}$ (with tilde). All random variables are expressed in capital letters, while the corresponding sample values are denoted in small letters: $m_{ij}, v_{ij}, \tilde{x}_{ij}^{(k)}, \tilde{m}_{ij}$, and $\tilde{v}_{ij}$, respectively.

First, we describe the conventional maximum likelihood estimation. This method estimates the optimal transformation parameters, $\hat{\boldsymbol{\theta}}$, using all observed values $x_{ij}^{(k)}$, while unknown parameters $\boldsymbol{\mu} \equiv \{\mu_{ij}\}$ are included in the estimation

$$\hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} \left[\max_{\boldsymbol{\mu}} \prod_{i,j,k} \mathcal{P}_{X_{ij}}(x_{ij}^{(k)}; \boldsymbol{\theta}, \mu_{ij}) \right], \tag{9}$$

where $\mathcal{P}_{\bullet}$ denotes the probability density function of random variable $\bullet$. Although $\mathcal{P}_Q$ is unknown, we can approximate $\mathcal{P}_{X_{ij}}$ by assuming that the transformed variable $\tilde{X}_{ij}$ follows normal distribution $\mathcal{N}(\tilde{\mu}_{ij}, 1)$:

$$\mathcal{P}_X(x) = \mathcal{P}_{\tilde{X}}(\tilde{x}) \left| \frac{\partial \tilde{x}}{\partial x} \right| \simeq \frac{1}{\sqrt{2\pi f_{\boldsymbol{\theta}}(x)}} \exp\left(-\frac{(\tilde{x}-\tilde{\mu})^2}{2} \right), \tag{10}$$

where we have omitted subscripts and superscripts for simplicity. Consequently, we get

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i,j} \left[-2n_{ij}\tilde{v}_{ij} - \sum_k \ln f_{\boldsymbol{\theta}}(x_{ij}^{(k)}) \right], \tag{11}$$

where $n_{ij} \equiv \frac{N_{ij}-1}{2}$. This method is substantially equivalent to the conventional method[4,5].

This method has a disadvantage; the likelihood of X_{ij} with small σ_{ij}^2 is always greater than that with standard σ_Z^2, although we expect the likelihood with standard σ_Z^2 to be highest. This results in high variances of the transformed data (as shown in Fig. 2(c)). Moreover, this method is unable to manage outliers using likelihood, although likelihood is a useful criterion for judging outliers as described in the next paragraph.

Our approach is more straightforward; we directly estimate the variances of the transformed data

$$\hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} \prod_{i,j} \mathcal{P}_{\tilde{V}_{ij}}(\tilde{v}_{ij}; N_{ij}, 1), \tag{12}$$

where $\mathcal{P}_{\tilde{V}_{ij}}$ is the probability density function of the unbiased variance of N_{ij} independent sample values from a normal distribution with variance $\sigma^2 = 1$, and is easily determined using the gamma distribution as

$$\mathcal{P}_{\tilde{V}}(\tilde{v}; N, \sigma^2) \equiv \frac{1}{\Gamma(n)} \left(\frac{n}{\sigma^2} \right)^n \tilde{v}^{n-1} \exp\left(\frac{n}{\sigma^2} \tilde{v} \right), \tag{13}$$

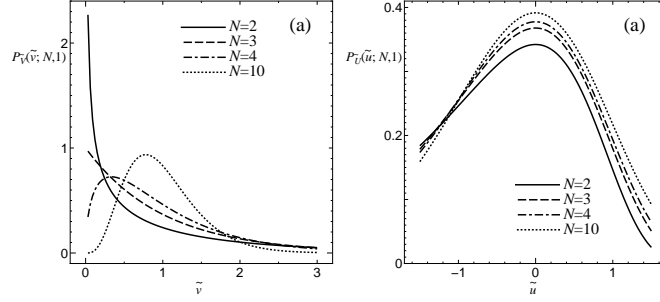where $n \equiv \frac{N-1}{2}$, while Γ is the gamma function.

Fig. 1. Probability density functions. (a) $\mathcal{P}_{\tilde{V}}(\tilde{v};N,1)$, (b) $\mathcal{P}_{\tilde{U}}(\tilde{u};N,0)$.

We can manage the effect of outliers (which mean extremely high variances, $\sigma_{ij}^2 \gg \sigma_Z^2$, or low variances, $\sigma_{ij}^2 \ll \sigma_Z^2$) by excluding the smallest $p\%$ of all the likelihoods, $\mathcal{P}_{\tilde{V}_{ij}}(\tilde{v}_{ij})$, from the estimation, where we predetermine the rate of outliers p (e.g. 10 to 20%, see the discussion for details) before the estimation. In the case of an iterative search for the optimal parameters, we exclude the outliers at every iterative step by sorting all $\mathcal{P}_{\tilde{V}_{ij}}(\tilde{v}_{ij})$. This exclusion of outliers according to likelihood enables the statistically appropriate management of outliers.

In the case of $N_{ij} = 1$, this method is also applicable; we replace $\tilde{V}_{ij}$ by $\tilde{V}_i$, which is the unbiased variance of the ith gene, and N_{ij} by the number of conditions. A slightly higher p may be better to use here because of the existence of differentially expressed genes.

Finally, we complete this estimation method by considering the case of small N_{ij} using a technique of information normalization. Although the case of small N_{ij} is commonplace, we encounter two problems with the method of Eq. (12): 1) the variances of the data transformed by this method tend to be smaller than 1 in the case of small N_{ij}, 2) this method is not applicable when $N_{ij} = 2$ or 3 because the probability density function of $\tilde{V}$ (Eq. (13)) has a single peak at $\tilde{v} = 0$ (Fig. 1(a)).

Specifically, instead of Eq. (12), we propose the following estimation:

$$\hat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}} \prod_{i,j} \mathcal{P}_{\tilde{U}_{ij}}(\tilde{u}_{ij};N_{ij},0), \tag{14}$$

$$\mathcal{P}_{\tilde{U}}(\tilde{u};N,\rho) \equiv \frac{n^{n-1/2}}{\Gamma(n)} \exp\left(n\left\{\frac{\tilde{u}-\rho}{\sqrt{n}} - \exp\frac{\tilde{u}-\rho}{\sqrt{n}}\right\}\right). \tag{15}$$

Here we use ρ instead of σ^2, and it is exactly determined by transforming σ^2 so that the Fisher information of ρ is always equal to 1 ($\mathrm{I}[\rho] = 1$);

$$\rho \equiv \int \sqrt{\mathrm{I}[\sigma^2]}d\sigma^2 = \sqrt{n}\ln\sigma^2, \tag{16}$$

$$\mathrm{I}[\sigma^2] \equiv \mathrm{E}\left[\left\{\frac{\partial}{\partial\sigma^2}\ln\mathcal{P}_{\tilde{V}}(\tilde{V};N,\sigma^2)\right\}^2\right] = \frac{n}{\sigma^4}. \tag{17}$$

This information normalization is somewhat analogous to variance stabilization (compare

6 *M. Inoue et al.*

Eqs. (5) and (16)) because the variance and Fisher information are reciprocal values at the Cramér-Rao bound. The maximum likelihood estimator $\tilde{U}$ of this parameter ρ is easily determined as

$$\tilde{U} = \sqrt{n}\ln\tilde{V}. \tag{18}$$

The probability density function of $\tilde{U}$ is given by Eq. (15) and shown in Fig. 1(b); the distribution always has a peak at $\tilde{u} = 0$ (corresponding to $\tilde{v} = 1$) regardless of N. This information normalization clearly improves the performance of variance stabilization. (Note that the maximum likelihood estimation is affected by the transformation of the applied stochastic variable(s) – $\tilde{V}_{ij}$ in this case – although it shows invariance against the transformation of parameters. Also note that we have defined $\tilde{V}$ as an unbiased estimator and $\tilde{U}$ as the maximum likelihood estimator for simplicity. Either estimator is applicable for $\tilde{V}$ and $\tilde{U}$; it does not affect the overall estimation.)

Equation (14) can be simplified as follows.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i,j} L_{ij}(\boldsymbol{\theta}), \tag{19}$$

$$L_{ij}(\boldsymbol{\theta}) \equiv \ln\frac{n_{ij}^{n_{ij}-1/2}}{\Gamma(n_{ij})} + n_{ij}(\ln\tilde{v}_{ij} - \tilde{v}_{ij}). \tag{20}$$

Moreover, in the case where all N_{ij} are the same value, a simpler form, $L_{ij}(\boldsymbol{\theta}) \equiv \ln\tilde{v}_{ij} - \tilde{v}_{ij}$, is available. To exclude outliers, we ignore the smallest $p\%$ of $L_{ij}(\boldsymbol{\theta})$.

4. Validation

We validated the proposed estimation using several experimental datasets. Here we show results from Affymetrix GeneChip U74A and Microarray Suite 5.0 software (Hubbell et al, 2002); the number of genes (or ESTs) was 12422, there were three conditions, and four observations were made for each condition ($N_{ij} = 4$). Before applying the transformation, we calibrated the data of each chip through multiplication so that the median value of every chip would be the same value.

Figure 2 shows scatter diagrams of the variance against the average. The data before transformation shows a strong dependence between the variances and the averages (Fig. 2(a)). This dependence is explained by the gene-expression microarray data model we have described; i.e., a quadratic relationship (gray line). This dependence remains even after the log-transformation (Fig. 2(b)), where the gray line denotes the moving average of variances.

The results from the conventional transformation using Eq. (11) (Fig. 2(c)) were not good; the moving average of the variances (gray line) clearly shows variance dependence on the average, and tends to be higher than 1, which may have been caused by outliers with small variances. On the other hand, the results from the proposed method (Fig. 2(d)) were good; the dependence of the variance on the average was almost completely eliminated, except for a small amount of the samples with small averages. Artificially generated ideal results are shown in Fig. 2(e) for reference, and these are roughly comparable with the results of the proposed transformation.
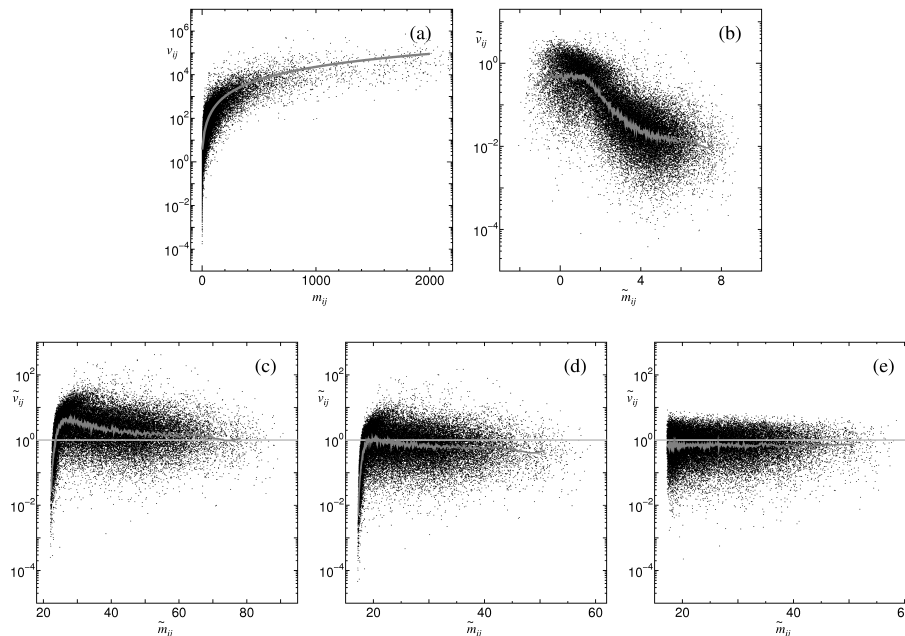
Fig. 2. Scatter diagrams of the variance against the average of each gene under each condition. (a) Raw data (after calibration). The gray line denotes quadratic function $f_{\hat{\theta}}$ determined by the proposed method. (b) Log-transformed data ($\tilde{x} = \ln x$). (c) Variance-stabilized data using the conventional method. (d) Variance-stabilized data using the proposed method (the rate of outliers p was set to 10%). (e) One of the ideal results, where the variances $\tilde{v}_{ij}$ were artificially generated from a normal distribution with unit variance. The averages $\tilde{m}_{ij}$ are exactly the same as those in (d). The gray lines in (b) - (e) each denote a moving average of the variances on a logarithmic scale.

5. Discussion

Information normalization provides several benefits. For example, from the viewpoint of Bayesian statistics, a uniform prior distribution of a normalized parameter will always satisfy Jeffrey's prior. In gradient descent learning, conventional gradient descent is always equivalent to natural gradient descent,[9] which is necessary to achieve Fisher efficiency. These benefits of information normalization may allow maximum likelihood estimation to produce appropriate results. However, information normalization of more than one parameter is not always possible; this requires elimination of Riemann-Christoffel curvature.

The rate of outliers, p, is the only parameter that we need to predetermine in the proposed method. The proposed method is more useful than one which cannot deal with outliers, but how to determine this outlier rate is a difficult problem. We empirically recommend setting this rate to a small but non-zero value (10 to 20%) to get better transformation parameters, because microarray data usually include a small number of outliers. Fortunately, the proposed method is rather robust and not greatly affected by this outlier rate. Of course, there should be better methods to determine p; for example, we can choose optimal p which gives the nearest empirical distribution to the model distribution in the sense of the

8   *M. Inoue et al.*

Kullback-Leibler divergence. We can also determine p as the rate of ESTs, or one based on it. We hope to develop these ideas more fully in our future work.

Further improvement of variance stabilization might be possible under a new gene-expression microarray data model. We emphasize that our framework is fairly general and may be applicable for a more elaborate relationship f, because we have not limited the class of f so far.

6. Conclusion

We have shown that the conventional method of estimating transformation parameters, which was proposed by Huber et al, has a disadvantage with respect to the management of outliers. We have developed a new estimation method that is able to exclude outliers efficiently even in the case of small sample size. The greater effectiveness of the proposed method, in comparison with the conventional method, was suggested through a demonstration based on experimental data.

References

1. D.M. Rocke and B. Durbin. "A model for measurement error for gene expression analysis," *J. Comput. Biol.* **8**, 557-569 (2001).

2. P. Munson. "A 'Consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations," *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data* (2001).

3. B. Durbin et al. "A variance-stabilizing transformation for gene-expression microarray data," *Bioinformatics*, **18** Suppl. 1, S105-S110 (2002).

4. W. Huber et al. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, **18** Suppl. 1, S96-104 (2002).

5. W. Huber et al. "Parameter estimation for the calibration and variance stabilization of microarray data," *Statistical Applications in Genetics and Molecular Biology*, **2**(1) Article 3, Berkeley Electronic Press (2003).

6. C.R. Rao. *Linear Statistical Inference and Its Applications, (2nd ed.).* John Wiley & Sons, New York, 1973.

7. D.M. Rocke, and B. Durbin. "Approximate variance-stabilizing transformations for gene-expression microarray data," *Bioinformatics*, **19**(8) 966-972 (2003).

8. E. Hubbell et. al. "Robust estimators for expression analysis," *Bioinformatics.* **18**(12), 1585-1592 (2002).

9. S. Amari. "Natural gradient works efficiently in learning," *Neural Comput.* **10**, 251-276 (1998).
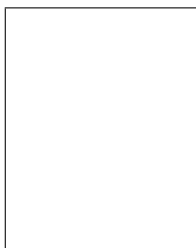
Masato Inoue, M.D. received his Ph.D. degree in Medicine from Kyoto University, Kyoto, Japan, in 2003. From April 2003 to March 2004, he was at Japanese Science and Technology Corporation or the RIKEN Brain Science Institute as a postdoctoral researcher. He is currently an assistant professor in the Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology.
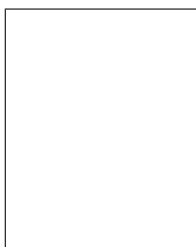
Shin-ichi Nishimura, M.D. Since April 2001 he has been a Ph.D. student at the Graduate School of Medicine and Faculty of Medicine, University of Tokyo. From April 2001 to March 2004, he was a Junior Research Associate at RIKEN BSI.
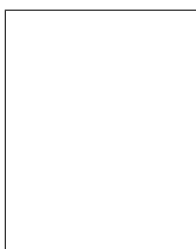
Gen Hori received his B.Sc. and M.Sc. degrees in mathematical engineering and his Ph.D. in information engineering in 1991, 1993, and 1996, respectively, all from the University of Tokyo. From April 1996 to September 1998, he was a Research Fellow of the Japan Society for the Promotion of Science. Since 1998, he has been a researcher with the Brain Science Institute, RIKEN, Japan.

His research interests include independent component analysis (ICA) and matrix dynamical systems applicable to signal processing.
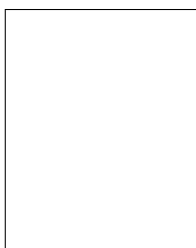
Hiroyuki Nakahara received his Ph.D. from the Dept. of General System Studies, University of Tokyo in 1997. Since then, he has worked at the RIKEN Brain Science Institute, currently as a researcher at the Laboratory for Mathematical Neuroscience. His research interests are computational neuroscience, bioinformatics, and machine learning.
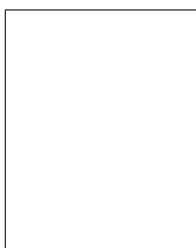
Michiko Saito received her M.Sc. degree from the Nara Institute of Science and Technology, Nara, Japan, in 1996. From Jan. 1999 to Dec. 2002, she worked for the Laboratory for Neurobiology for Synapse at the RIKEN Brain Science Institute as a member of the technical staff. She is currently a researcher in the Animal Cell Function, Research and Education Center for Genetic Information, Nara Institute of Science and Technology.
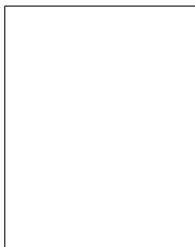
Yoshihiro Yoshihara is the Laboratory Head for the Laboratory for Neurobiology of Synapse at the RIKEN Brain Science Institute. He received his Ph.D. in Pharmaceutical Sciences from Kyoto University in 1989. His research focuses on molecular and cellular mechanisms underlying the development and functions of the olfactory system.

Shun-ichi Amari received his Doctorate of Engineering degree from the University of Tokyo in 1963. He was an Associate and then Full Professor at the Department of Mathematical Engineering and Information Physics, the University of Tokyo, and is now Professor-Emeritus at the University of Tokyo. He is the Director of the RIKEN Brain Science Institute, Saitama, Japan.

He has been engaged in research in wide ranging areas of mathematical engineering and applied mathematics, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, mathematical foundations of neural networks, and information geometry.

Dr. Amari served as President of the International Neural Network Society, Council member of the Bernoulli Society for Mathematical Statistics and Probability Theory, and Vice President of the Institute of Electrical, Information and Communication Engineers. He was founding Co-Editor-in-Chief of Neural Networks. He has been awarded the Japan Academy Award, IEEE Neural Networks Pioneer Award, IEEE Emanuel R. Piore Award, Neurocomputing best paper award, IEEE Signal Processing Society best award, and NEC C&C Prize, among many others.